LONDON SCHOOL OF ECONOMICS

AND POLITICAL SCIENCE

Information Organization of Social Media Platforms

The Case of Last.fm

Akarapat Charoenpanich

A thesis submitted to the Department of Management of the London School of Economics and Political Science for the degree of Doctor of Philosophy, London, September 2017

Information Organization of Social Media Platforms: The Case of Last.fm

Abstract

Social media has become a global phenomenon. Currently, there are 2 billion active users on Facebook. However, much of the research on social media is about the consumption side of social media rather than the production or operational aspects of social media. Although research on the production side is still relatively small, it is growing, indicating that it is a fruitful area to study. This thesis attempts to contribute to this area of research to unravel the inner operations of social media with one key research question: ***How does social media platform organize information?*** The theory of digital object of Kallinikos et al. (2013) is used to investigate this question. Information display that users of a social media platform interact with is a digital object and it is constructed by two key components which are a database and algorithms. The database and the algorithms shape how information is being organized on information displays, and these influence user behaviors which are then captured as social data in the database

This thesis also critically examines the technology of recommender system by importing engineering literature on information filtering and retrieval. While newsfeed algorithm such as EdgeRank of Facebook has already been critically examined, information systems and media scholars have yet to investigate recommendation algorithms, despite the fact that they have been widely deployed all over the Internet. It is found that the key weakness of recommendation algorithms is their inability to recommend novel items. This is because the main tenet of any recommender system is to "recommend similar items to those that users already like". Fortunately, this problem can be alleviated when recommender system is being deployed in the digital information environment of social media platforms.

In turn, seven theoretical conjectures can be postulated. These are (1) navigation of information display as assembled by social media is highly interactive, (2) information organization of social media is highly unstable which would also render user behaviors unstable, (3) quality of data aggregation casts significant implications on user behaviors, (4) the amount of data captured by social media platforms limits the usefulness of their information displays, (5) output from the recommendation algorithm (recommendation list) casts real implications on user behaviors, (6) circle of friends on a social network can influence user behaviors, and (7) metadata attached to items being displayed casts influence on user behaviors. Data from Last.fm, a social media for music discovery, is used to evaluate these conjectures. The analysis supported most of the conjectures except the instability of information display and the importance of metadata attached to items being displayed. Some kinds of information organization are more stable than initially expected and some kinds of user generated contents are not so important for user behaviors.

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent.

I warrant that this authorization does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 71,359 words

Statement of conjoint work

I confirm that the third paper was jointly co-authored with Aleksi Aaltonen (Warwick Business School) and I contributed 50% of this written work.

I confirm that the fourth paper was jointly co-authored with Cristina Alaimo (University of Surrey) and I contributed 25% of this written work.

Statement of use of third party for editorial help

I confirm that my thesis was copy-edited throughout (except for the third and the fourth papers), by proofreading service provided by http://fastwork.co for conventions of language, spelling and grammar.

Acknowledgements

First of all, I would like to extend my utmost gratitude to my supervisor, Professor Jannis Kallinikos, for his valuable advices, his continued support and his patience with me throughout these years.

I would like to say thank you to Aleksi Aaltonen and Cristina Alaimo for co-authoring papers with me. I have learnt a lot from them during the process.

I would like to say thank you to all participants at the Information System and Innovation Group (ISIG) seminars at the LSE and the European Conference of Information Systems 2015 (ECIS 2015) as their valuable advices have greatly shaped my work.

Also, I would like to say thank you to participants on websites such as Stackoverflow.com. Without them, I would not have learnt to write my own code to collect my own data from the Internet.

Furthermore, I would like to say thank you to P Nok, who has always been my mentor, my ex-colleagues and friends. Without them, I may not have chosen to pursue for this PhD degree in the first place.

Finally, I give my deepest and lifelong thankfulness to my parents and my family members for always being there for me. I cherish their unconditional love and supports and will always see this as the greatest gift ever given to me.

Original papers

The thesis includes the following papers:

1. Charoenpanich, Akarapat (2017). "Literature Review of the Production Side of Social Media"
2. Charoenpanich, Akarapat (2017). "Ranking and Information Display in Social Media Platforms"
3. Charoenpanich, Akarapat and Aaltonen, Aleksi (2015). "(How) Does Data-based Music Discovery Work?" In *European Conference on Information System 2015 (ECIS 2015)*
4. Alaimo, Cristina and Charoenpanich, Akarapat (2017). "More than Networks: Social Media as Infrastructures" *Manuscript under revision for subsequent resubmission to MISQ*

All of the work has been carried out following my registration in the PhD in Information Systems program in the Information Systems and Innovation Group, Department of Management, LSE.

Table of Contents

The following images have been redacted.

Page 49, Figure 7
Page 59, Figure 9
Page 121, Figure 1 & 2
Page 136, Figure 4
Page 137, Figure 5
Page 138, Figure 6
Page 140, Figure 7

Introduction

Social media has now become a global phenomenon with the staggering 2 billion active users on Facebook representing a quarter of the global population. A widely cited article by Kaplan and Haenlein (2010) defines social media as "a group of Internet based applications that build on the ideological and technological foundation of Web 2.0 and allow the creation and exchange of user generated content". Six categories of social media are blogs, collaborative projects (e.g. Wikipedia), social networking sites (e.g. Facebook), content communities (e.g. YouTube), virtual social worlds (e.g. Second Life) and virtual game worlds (e.g. World of Warcraft). Most research relates to the consumption side of social media; however, this thesis studies the production side or the operational aspects, aims to unravel the inner operations of social media and asks the following question: ***How does the social media platform organize information?*** Particular attention is paid to the construction and implication of recommender systems and their by-products (similarity networks) in the social media environment. This is a fruitful area of study because research regarding the operational aspects is still relatively scant and recommender systems have yet to be subjected to critical scrutiny like search engine algorithms (Introna and Nissenbaum, 2000) and the EdgeRank algorithm of Facebook (Bucher, 2012). Recommender systems have been deployed all over the Internet to recommend things users might like or want to do. Arguably, they are one of the most powerful algorithms which shape user online behavior yet critical scrutiny is lacking.

The thesis consists of a cover paper and four individual papers (Figure 1). The cover paper includes seven sections. The first discusses the distinction between the consumption and production side or operational aspects of social media research and demonstrates that the operational aspects are a fruitful, small but growing study area. Furthermore, it is imperative to understand how the workings of the inner operations of social media orchestrate the

behaviors of users. The second section presents a theoretical framework of information organization on social media platforms whereby interfaces are constructed from data and algorithms to display information and shape user behavior. It synthesizes the technical literature on information retrieval and filtering and examines the organization theory of Weinberger (2008) and the theory of digital objects (Kallinikos et al., 2013). The third and fourth sections discuss research methodology and the empirical objective of this thesis as Last.fm, a social media platform for music discovery. Social big data is captured and utilized to find the statistical association between different components underlying the inner operations of social media platforms and user behaviors. If such statistical association is detected, then it implies that those components matter for the operation of Last.fm, validating or falsifying conjectures derived from the theoretical framework postulated in the second section. The fifth section presents research procedures and contributions, including summaries of the four papers and a description of how social media organize information, while the sixth identifies further research areas.

Figure 1: The thesis papers

The thesis papers	
Paper 1	Charoenpanich, Akarapat (2017). "Literature Review of the Production Side of Social Media"
Paper 2	Charoenpanich, Akarapat (2017) "Ranking and Information Display in Social Media Platforms"
Paper 3	Charoenpanich, Akarapat and Aaltonen, Aleksi (2015). "(How) Does Data-based Music Discovery Work?" In *European Conference on Information System 2015 (ECIS 2015)*
Paper 4	Alaimo, Cristina and Charoenpanich, Akarapat (2017) "More than Networks: Social Media as Infrastructures" *Manuscript under revision for subsequent resubmission to MISQ*

Social media research on the consumption side vs. the production side

Perhaps, one of the most influential reviews of social media is by boyd and Ellison (2007) in their paper entitled "Social Network Sites: Definition, History, and Scholarship". Arguably, this paper plays an important part in establishing the field of study on social media. They define social networking sites, a kind of social media, as merely "web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection and (3) view and traverse their list of connections and those made by others within the system". It is interesting that boyd and Ellison (*ibid.*) distinguish social network sites from social networking sites and view the latter as a subset of the former. They suggest that networking (meeting with strangers) is not the primary practice of many social media users who merely communicate

with people who are already a part of their offline social network. This demonstrates the user-oriented nature of their research agenda as social network sites are merely there to support already existing offline relationships. In general, the inner operation of social network sites does not enable individuals to forge new connections between themselves and does not interfere with user behavior. Further, they discuss four research agenda: (1) impression management and friendship performance, (2) network and network structure, (3) bridging online and offline networks and (4) privacy. All these can be considered as studies on the consumption side of social media or studies which focus on the uses of social media rather than its inner operation.

Kane et al. (2014) modify the meaning of social networking sites as proposed by boyd and Ellison (2007) and highlight four features shared by social media technologies as digital profile, search and privacy, relational ties and network transparency. While the latter two features are similar to the definition of boyd and Ellison (*ibid.*), the former two are not. Nowadays, digital profiles extend beyond exclusive intentional and conscious construction by users toward incorporating automatic and passive records of user activity. People can also access content on social networking sites without directly viewing digital profiles. For example, content streams can be automatically filtered and users might engage in activities such as searching for keywords in LinkedIn profiles to find people with particular skills or experience. The ability to search for contents has raised concerns about data protection; therefore, privacy has become a more significant social media issue. Most social media sites provide control features for users to specify who can access the content they contribute. Kane et al. (2014) attempt unsuccessfully to discuss the inner workings of social media. While they question the idea of a bounded system embedded in the definition of boyd and Ellison (2007) they do not incorporate this into their own new definition. Therefore, some argue that their paper still focuses on the consumption side of social media.

Zhang and Leung (2015) review the literature on social networking services in the top six communication journals between 2006 and 2011. They discover that most of the papers fit into the four themes as discussed by boyd and Ellison (2007). In addition, many papers demonstrate how trust, attraction, emotional closeness, emotional support and perceived social support are facilitated by the use of social networking services, while several studies adopt a psychological approach and incorporate intrapersonal psychological traits such as self-esteem, collective self-esteem, happiness, satisfaction, emotional openness and extraversion. In general, past research reports that people experience more happiness and excitement when using social networking sites. Nonetheless, some personality characteristics negatively affect individual's offline and online communication such as loneliness, jealousy, communication apprehension, narcissism and neuroticism. Zhang and Leung (*ibid.*) point out that future research should emphasize the role of networks, improve measures of use, rethink the nature of relationship and friendship on social networking sites, consider the dynamic adoption process and expand to cross-contextual and cross-cultural contexts.

Rains and Brunner (2014) review research related to social networking services published in six interdisciplinary journals between 1997 and 2003. They determine that over two-thirds of the studies are explicitly limited to a single company and that Facebook is examined by approximately 80% of authors. Reviews of microblogs are even more concentrated with over 90% in the six journals limited solely to Twitter. These findings concur with Zhang and Leung (2015) and Zhang et al. (2015a) who point out the importance of Facebook. Zhang et al. (*ibid.*) search for papers related to social media on the Citation Database for the research areas of "business and economics" and "computer science". They conclude that studies on social media in different disciplines are not well combined and attempt to obtain a better picture by choosing to investigate the discipline of management and computer science. On the one hand, their data purport that the most cited study is a paper by Kaplan and Haenlein

(2010) entitled “Users of the World Unite! The Challenges and Opportunities of Social Media” which makes the study of social media more explicit and systematic. On the other hand, the paper enjoying the highest centrality in the bibliography networks is the classic study by Granovetter (1973) entitled “The Strength of Weak Ties”. The most cited papers on the business side concern word of mouth communication, while the most cited in the field of computer science discusses analytical techniques (topic modeling and social network analysis). Finally, they determine that studies of social media increase rapidly after 2009 when Facebook emerges as one of the top keywords.

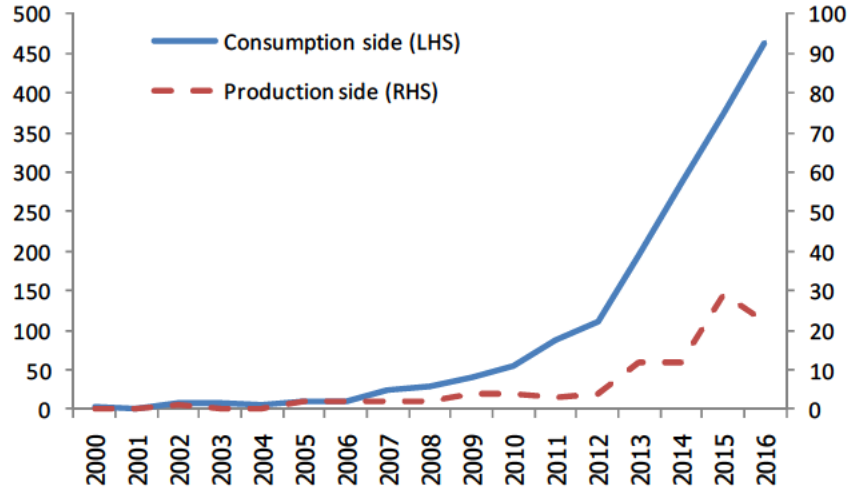
So, how did previous authors study Facebook? Caers et al. (2013) examine scientific peer reviewed articles on Facebook between 2006 and 2012 from the ISI Web of Knowledge. Thematic topics covered in their corpus include initial motivations to join Facebook, characteristics of Facebook users, building and maintaining a Facebook network, motive for disclosing information on Facebook and the effect of disclosing information on Facebook. They also review papers on the organizational aspects of Facebook including, how it reaches out to customers and future staff. In addition, they identify numerous gaps in the current studies of Facebook, including why former users decide to abandon the site, cyber-bullying and the extent to which information disclosed by users reflects their actual personality traits, motivation and competence. Wilson et al. (2012) sort papers on Facebook into five categories as descriptive analysis of users, motivations for using Facebook, identity presentation, the role of Facebook in social interactions and privacy and information disclosure.

What is apparent from the reviews above is that none focus on the inner workings of the operational aspects of social media and how they, in turn, shape users behaviors. It is interesting that this is not even considered as a fruitful area of future research by authors who focus on the consumption side of social media. Research on the operational aspects of social media is clearly lacking and this thesis intends to fill this research gap and unravel how social

media constructs and displays information. The first paper is entitled "Literature Review of the Production Side of Social Media" and analyses papers produced from eight journals in the field of information systems (Senior Scholars' Basket of Journals) and 10 journals from the field of media. In total, 1,732 papers are classified on the consumption side of social media while only 99 are classified on the operational aspects or the production side. The former concern why social media is used, with particular linkage to psychological traits and motivations and emphasis on diverse usages and both utopian and dystopian consequences. Figure 2 plots the number of papers relating to the consumption side and the production side of social media through time. On the one hand, papers published on the consumption side start to gain traction by 2007 and accelerate faster in 2013. On the other hand, research on the production side grows steadily from 2013. This is a relatively small but growing and important area for research and may prove to be particularly fruitful.

Figure 2: Number of papers on the consumption side and operational aspects of social media

Unit: Paper count



Theoretical framework: How do social media platforms organize information?

The overarching principle of social media sites is simple; they collect data from their users and organize that social data to construct interfaces and information displays to maximize user engagement and profitability. This thesis aims to unravel the organization of these interfaces and information displays from the software perspective as data and algorithms are the input into the construction of information displays. The interfaces are not neutral. Grosser (2014) points out that "despite a common belief that networks and computational systems are neutral actors enabling human communication and creativity, these systems enact a series of constraints on those who use them, directing their actions, limiting their options and constructing them as users".

Data are fundamental for the operation of social media. According to van Dijck and Poell (2013), social media logic is well grounded in the condition of datatification which can be referred to as the ability to "render into data many aspects of the world that have never been quantified before". Likewise, Kallinikos and Constantiou (2015b) cite Gillespie (2014) who argues that "algorithms are inert, meaningless machines until paired with databases upon which to function. A sociological inquiry into an algorithm must always grapple with the databases to which it is wedded". Therefore, this section begins by discussing data collected by social media platforms and then examines algorithms and interfaces. Next, the theory of Kallinikos et al. (2013) is discussed as data, algorithm and interface are all digital objects. Finally, theoretical syntheses and numerous conjectures derived from the theory are presented.

Data

Social media collect so-called 'social data' from users. Alaimo and Kallinikos (2016) identify three kinds of social data as profile data, behavioral data and user generated content (UGC). They point out that behavioral data are the most valuable social data. Social media assembles standardized activities for users including 'tagging', 'liking' and 'following'. As users enact these pre-programmed standardized activities, behavioral data are generated. These behavioral data sets are 'discrete and granular' since their creation entails drastic simplification as compared to action in everyday life outside the social media platforms. Hence, they become countable and can be further processed more easily. This is a virtue of behavioral data. Take 'liking' as an example. Facebook regards every 'like' as being the same but 'like' may have a different meaning for different people in actual everyday life. This is because of its codification which entails a simple premise that everything (e.g. users, comments, photos) can be seen as objects and every object can be connected together by pre-programmed standardized activities such as 'tagging', 'liking' and 'following'.

Behavioral data is also useful because it adds value to profile data and UGC. Profile data entails descriptive data about individuals and this gains value when it is mapped with behavioral data. On the other hand, UGC entails huge amounts of unstructured data seldom used directly. However, once combined with behavioral data UGC becomes computable. For example, 'tagging' of photos in Flickr allows users to navigate through a huge database of images. This is another virtue of behavioral data.

Constantiou and Kallinikos (2015b) note that "data generation is lifted out of the prevailing expert-dominated cultures by which the information needs of practice fields have been defined and data collected stored". This is "the outcome of the fundamental fact of making online interaction and the activities of large, shifting, heterogeneous and dispersed populations of users (mostly lay people) the drivers and carriers of data generation".

Similarly, Pletrobruno (2013) analyses the transmission of intangible UNESCO heritage videos on YouTube and determines that contributions of lay people countered the official heritage narrative, while Zervas and Sampson (2014) analyses the implications of tagging digital educational resources and suggest that tagging by lay people can enlarge relevant metadata as compared to tagging performed exclusively by experts. Kallinikos and Tempini (2014) point out that the kind of data collection by PatientsLikeMe, a social media platform for patients, differs to how data is collected for medical research as it is self-reported and does not rely on clinical interviews performed in institutional hospital environments by doctors and nurses (Kallinikos and Tempini, 2014).

Constantiou and Kallinikos (2015a) identify four characteristics of social data. First, it is heterogeneous and often useful when amalgamated in an aggregation which uncovers context generality rather than a specific and contingent character of each data point. Second, it escapes the systematic nature of professional classification. Third, it crosses the border of alphanumerical systems and includes varying cultural artifacts cast in the media of text, image and sound while finally, it requires constant renewal and updating.

Helmond (2015) notes that social media sites are transformed into social media platforms when they establish application programming interfaces (APIs) which render the platforms reprogrammable by third parties developers. APIs were initially implemented as business-to-business (B2B) solutions for e-commerce, enabling transactions and sales management. For example, Salesforce established APIs in 1999, eBay in 2001 and Amazon in 2002. In the mid-2000s, social media sites started to establish their own APIs. Delicious established APIs in 2003, Flickr in 2004 and Last.fm, Facebook and Twitter in 2006. Developers can access platforms' data and functionality through API's enabling them to read, write and delete user data. Dissemination of the so-called widgets as plugin modular components enables integration of platforms' content and functionality into another website using a few lines of

code. This includes social plugins such as the 'like' button developed by Facebook. Technically, these social buttons function as API calls and send specific requests to Facebook's platform, for example, to ascertain the number of people who like the post or to publish the likes on the user's timeline. APIs can change as the business models of social media change. In the past, Twitter had a reputation as a data accessible platform since the Twitter API allowed easy scrape or download of massive amounts of data. However, Twitter imposed a download restriction of only 1% of traffic in 2011 and encouraged users to purchase data through a Twitter reseller such as Gnip (Felt, 2016).

According to Gerlitz and Helmond (2013), this has resulted in the 'like economy'. Facebook eventually extended beyond the limit of its platform and offered widgets which can turn websites and applications into a part of its platform. Social graph is an important component of Facebook as the representation of people and their connections to other people as well as objects within the platform. In April 2010, Facebook launched Open Graph Protocol, which allows external websites and applications to be integrated with Facebook's Social Graph. Currently, more than 7 million applications and websites are integrated with the platform. Social plugin allows users to engage with content outside the platform through Facebook based activities such as liking, sharing and commenting. Once users click on the like or share button attached to external contents, these then become available for further liking and comment within the Facebook platform, generating additional data flow back to the external counter. Furthermore, data flows back to webmasters in the form of Facebook Insights with, for example, reports on the basic demographics of likers such as age, gender and location. Hence, webmasters are happy to grant Facebook real estate on their web pages in exchange for user engagement and Facebook Insights. Applications can also be integrated with Facebook's Open Graph. Using an eReading device called Kobo as an example, Facebook registers when users start reading a book on the device and will inform the user's friends on

their newsfeeds when this occurs. Chains of comment/like then follow these announcements which can be tracked by Kobo staff (Kaldrack and Rohle, 2014). In the same vein as webmasters, third party developers are happy to integrate their applications with Facebook in exchange for user engagements and insights.

This attempt by social media to extend their operations throughout the Internet has given rise to a data management problem which is identified by Tempini (2015). He analyses PatientsLikeMe, a social media platform for patients as mentioned above, whereby users can upload and track their medical conditions (diseases, symptoms and treatments) and find patients with similar conditions who can offer support by sharing their own experiences. Self-reported data are then exploited for scientific and commercial medical research. PatientsLikeMe has produced 37 scientific publications based on data contributed by more than 220,000 patients. Tempini (2005) discusses the data management challenges faced by PatientsLikeMe with conflicting demand for local context flexibility and data specificity richness. All patients differ from each other (for example, they might have different levels of medical literacy) and each needs to be treated as an individual to enhance engagement on PatientsLikeMe. For example, patients can even request the creation of new medical entities or definitions that are not available in the database. However, data created by users may not be deemed specific enough for medical research and local context flexibility may be limited to enhance data specificity. It might be necessary to differentiate between patients suffering from taxonomically close conditions (subtypes of the same parent condition). In this case, PatientsLikeMe may allow users to input only subtypes in the system, but not the parent condition. While this increases data specificity richness, some patients may not recognize the subtypes and overall engagement is dampened.

Fundamentally, this is a problem of data aggregation as things to be counted (disease in the case of PatientsLikeMe) are being named differently by different users. This problem

emerges because data collection is no longer in the hands of experts (Constantiou and Kallinikos, 2015b). Today, users are given the freedom to name objects in their everyday lives. The problem becomes more severe when social media sites transform into social media platforms and connect with the wider digital ecosystem. The mechanism as discussed by Tempini (2015) to cope with the problem no longer works as it is only site-specific and cannot be applied to the whole digital ecosystem. It is not difficult to imagine that the 'like economy' of Facebook must be experiencing the same problem as the same object is likely to be assigned with different identifiers by different webmasters. This makes data aggregation performed by Facebook dubious and incomplete; it can no longer combine data across the same object with different identifiers. This is very important and can potentially harm components of Facebook that rely upon such data aggregation as input, such as recommender systems. Users can wield social data on social media to construct their own identity because of 'persistent labeling', whereby each user is assigned with a permanent identifier. Together with deep profiling or availability of past interaction archives, users can learn about the identity of other users (Ma and Agarwal, 2007). However, this 'persistent labeling' is not available for objects which are being counted and processed by social media platforms.

Distributed labeling is discussed in depth by Parsons and Wiersma (2014). They suggest that users may name an object differently because of diverse levels of expertise. Laymen often accurately classify an object at 'basic level' which is widely accepted in cognitive psychology as the generally preferred classification level of non-experts. This is an intermediate taxonomy level (for example bird is a level higher than American Robin but a level lower than animal) and is often the first class people think of when they encounter an instance. On the other hand, experts are likely to classify an object with more specificity. Equally troubling is that even if all users follow the 'basic level' category to label objects correctly there might still be variation in names because of existing synonyms, spelling errors

or added symbols (as brackets, tildes) that do not carry additional meaning. While humans can easily recognize and resolve these variations, computers cannot. Gehl (2011) points out that social media sometimes utilize their users as information processors. For example, Digg users sift through massive amounts of digital information and rate them, allowing Digg to sort and organize digital information on the whole Internet.

Weinberger (2008) correctly realizes that objects must have handholds so that information can be coalesced around them. This may include things like the Universal Product Code (UPC) for trade items and International Standard Book Number (ISBN) for books. Essentially, if something cannot be pinned down, then information cannot be coalesced around it. This can be a problem in the digital ecosystem as there is no centralized authority to produce and maintain these handholds because the amount of data in circulation far outstrips the volume of data that can be effectively managed by a single organization. According to Weinberger (*ibid.*), standardized identifiers will break down, opening up the opportunity for users to label things themselves. For example, image collection hosted by Flickr at 100+ million completely outstrips that of professionally managed archives like Bateman and Corbis. Therefore, giving up control may be the only solution to manage massive amounts of content generated in the digital ecosystem. Thus, the same labels might be applied to different items and different items may have the same labels, depending on who talks about them. This complexity requires effective management to organize information in social media platforms.

Algorithm

"Social media platforms don't just guide distort and facilitate social activity, they also delete some of it. They don't just link users together; they also suspend them. They don't just circulate our images and posts, they also algorithmically promote some over others. Platforms pick and choose."

Gillespie (2015)

So, "platforms pick and choose" according to Gillespie (2015) and the underlying algorithm appears invisible. Beer (2009) notes that "software [is] 'sinking' into and 'sorting' aspects of our everyday lives" and cites Thrift (2005) who suggests that "software has come to intervene in nearly all aspects of everyday life and has begun to sink into its taken-for-granted background". Other scholars also express similar concerns. Baym (2015) points out that opaque algorithms are filtering what one sees; users can neither understand nor influence these filtering mechanisms or comprehend the interest they serve. Sandvig (2015) is concerned about the secret process that determines relevance, judging whether something will be shown at all. For example, Facebook evaluates user content and may decide not to show some posts to anyone. Braun (2015) points out that mechanical editors exist on Facebook who decide algorithmically which posts and topics warrant inclusion into the continuous and often overwhelming feed of information delivered to users. Bucher (2015) comments that users do not simply article and make their network visible as networks are also articulated and made visible for them by underlying software and algorithms. Lastly, Shah (2015) considers that most information is communicated in social media between machine and machine and not consumed by humans. Along similar lines, Beer (2009) cites Hayles (2006) who claims that in "highly developed and networked societies … human awareness comprises the tip of a huge pyramid of data flows, most of which occur between machines".

van Dijck and Poell (2013) deconstruct social media logic and determine that algorithms can steer users' contributions and shape all kinds of activities such as liking, favoriting, recommending and sharing. This has culminated into automated connectivity which connects users to content, users to users, platforms to users, users to advertisers and platforms to platforms. For example, Facebook and Linkedin present users with lists of 'people you may know', Flickr presents users with 'groups you may be interested in' and Amazon recommends items as 'people who bought this item also bought'. Compared to mass media, algorithmic assessment of information has replaced reliance upon credential experts and scientific evidence. Nowadays, algorithms have the ability to boost the popularity of people, things or ideas. Facebook's EdgeRank and Twitter's Trending Topics have the ability to promote some material over others and each social media creates its own popularity metrics and tries to make them meaningful in social life offline. This includes view statistics for YouTube, friend statistics for Facebook and follower counts for Twitter.

Social media is filled with metrics. Grosser (2014) defines metrics as "enumeration of data categories or groups that are easily obtained via typical database operations and represent a measurement of that data". Metrics rely upon perhaps the most basic algorithm (summation) and are arguably the basic blocks of more complex algorithms such as Facebook's EdgeRank targeting advertisements, numerous recommendations and matching systems. Facebook is filled with metrics such as numbers of likes, comments, shares, friends, mutual friends, pending notifications, events, friend requests, message waiting, chats waiting, photos, places and much more. Facebook also produces metrics, which are hidden from users, for example how many objects users like per hour, how many advertisements they click and the effectiveness of the list of 'people you may know' at getting users to add more friends. The question is, how does Facebook choose which metrics to reveal to its users? The primary criterion for making such a decision is whether a particular metric will increase or decrease

user participation. For example, users are more likely to click on an advertisement if they see that a lot of people have already liked the object. Indeed, Facebook has the status of perpetual beta, whereby hundreds of experiments on small design variations and features are rolled out every day. Impacts of alternative designs are compared and the most efficient at fostering user participation are selected. Users are unaware that they are the subjects of these tests and they have no choice but to steer toward the most efficient designs (Heyman and Pierson, 2015).

Two algorithms as 'newsfeed' and 'targeted advertisement' are extensively deployed by social media platforms. Bucher (2012) and Birkbak and Carisen (2016) discuss Facebook's EdgeRank algorithm which underlies the construction of newsfeed. Bucher (2012) points out that the EdgeRank algorithm leads to a 'treat of invisibility' whereby users are encouraged to participate on social media platforms. For targeted advertisement, Heyman and Pierson (2015) identify the existence of paid solutions to the 'treat of invisibility' whereby users can pay social media platforms to get their content promoted. An example is Sponsored Story (SPS) that Facebook uses to stimulate business whereby posts of friends which are related to specific pages, applications or other items that advertisers want to promote achieve higher ranking by the EdgeRank algorithm and so are more likely to appear in the newsfeeds of users targeted by advertising according to their interests or profile details. Although Facebook no longer provides SPS which was replaced with separate advertising services in 2014, the basic idea behind SPS remains.

'Recommendation' is another algorithm extensively deployed by social media platforms which has not been subjected to critical scrutiny. Although some papers exist on recommender systems, for example Colace et al. (2015) their goal is to introduce new kinds of recommendation algorithms rather than a critical examination. This thesis pays particular attention to the construction, implication and technology of recommender systems, building

on the literature concerning data retrieval and filtering which has yet to be not imported into information systems and media analysis. The information organization theory of Weinberger (2008) is also discussed as this provides a broader perspective on the construction of interfaces, information displays and information organization on social media.

Information filtering (IF), information retrieval (IR) and recommendation algorithms

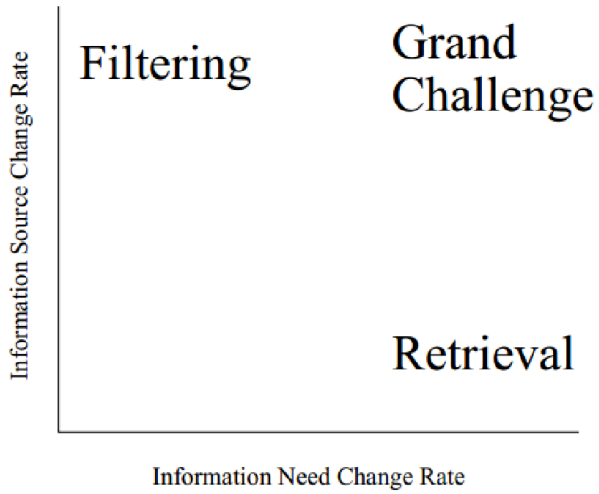
IF and IR

IF and IR are two related types of information seeking; a process whereby users obtain data from automated systems (Marchionini, 1995). Development of the two systems began at roughly the same time during the 1940s-1950s (Bush, 1945; Holmstrom, 1948; Luhn, 1958) in response to electronic information overload and the increasing availability of computing power which continues to motivate development and deployment of IF/IR systems today (Oard, 1997; Birkbak and Carlsen, 2016). While IF systems can be seen as a subset of IR systems (with "zero query" search), recommender systems are very much a subset of IF systems. There are also other types of IF systems such as personal email filters and newsgroup filters (Hanani et al., 2000). The following section compares and contrasts IF and IR systems and discusses each in turn.

Belkin and Croft (1992) point out that IF and IR systems can be viewed as different sides of the same coin. IF systems select relevant documents according to the long-term interests of users, while IR systems select relevant documents according to one-off queries. Information sources of the latter are very much static, while the former is a constant influx of new documents (Belkin and Croft, 1992; Riordan and Sorensen, 1997). Oard (1997) formulates a problem space characterized by information need change rate and information source change rate as depicted in Figure 3 whereby IF/IR systems occupy different areas. He notes that the

grand challenge for information detection systems is to match rapidly changing information with highly variable interests and neither IF nor IR systems generally accomplish this.

Figure 3: IF and IR problem space



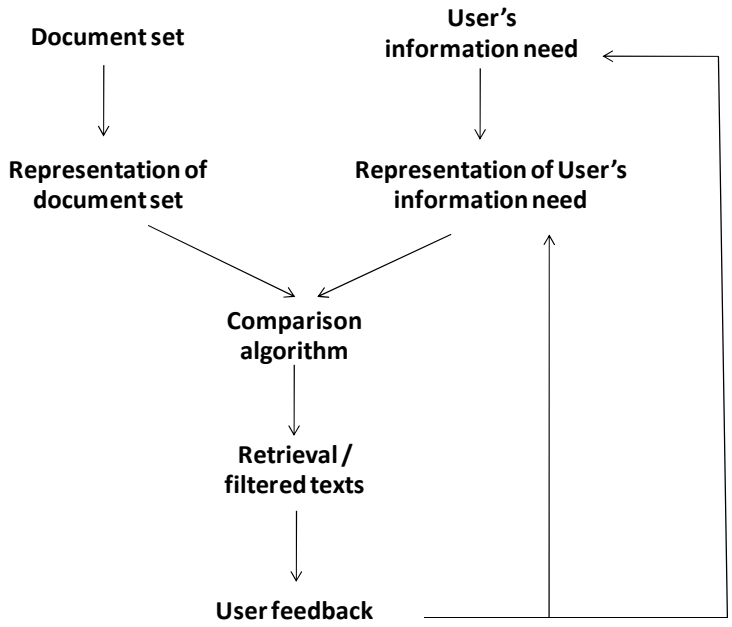
Source: Oard (1997)

The architectures of IF/IR systems are very similar (Belkin and Croft, 1992; Riordan and Sorensen, 1997) with both composed of three components. First is representation whereby user's information need (query or profile) and document set (fixed or dynamic) must be represented in a manner so that comparisons can be made between the two. Initially, IF/IR systems were built to filter/retrieve text. To represent text in computer readable formats they are often broken down into terms and then they undergo pre-processing such as stemming and removal of stop words (Porter, 1980). Numeric weights are then assigned to each term to measure how effective each is at distinguishing a document from other documents (Greengrass, 2000).

Figure 4 reveals how the different components of IF/IR systems relate to one another (Riordan and Sorensen, 1997). Once representations of documents and information needs are extracted, they are compared to one another according to a comparison algorithm. Relevant documents are then retrieved/filtered and users provide feedbacks of the success. In turn,

these feedbacks alter the representations of information needs or the information needs themselves. There might be better queries that serve the information needs or the information needs themselves might change after users learn from the retrieved/filtered documents.

Figure 4: Architecture of IF/IR systems



Source: Riordan and Sorensen (1997)

Oard (1997) explains this process another way as how IF/IR systems entail three steps of collection, detection and display. Detection happens when the representation of documents and information needs are compared according to a comparison algorithm. The resulting documents are then displayed at the interface for users to provide feedbacks. The three steps are related to one another. For example, if the aim of the display is to rank output, then the comparison algorithm must be able to provide some basis (e.g. a numeric "status value") from which ranking can be constructed.

There are three key classical IR techniques: (1) exact match and Boolean model, (2) vector space model and (3) probabilistic model (Baeza-Yates and Ribeiro-Neto, 2011). Exact match and Boolean model are considered to be the weakest classical methods as they do not rank

documents according to relevancy. However, ranking documents is superior to presenting users with a set of documents as it allows humans and machines to synergistically achieve better performance than either can attain alone and enhances user satisfaction (Oard, 1997; Winiwater et al., 1997). To rank documents, comparison algorithms must attach numerical status value to each document in a vector space model or a probabilistic model. However, there is controversy on whether a probabilistic model outperforms a vector space model (Croft and Harper, 1979; Salton and Buckley, 1988) with the dominant thought among researchers and practitioners of IR that the vector space model is better (Baeza-Yates and Ribeiro-Neto, 2011).

IF can be considered as a "zero query" approach to IR such that user profiles are maintained and user models are constructed to predict user preferences. This allows documents to be filtered automatically. There are two key paradigms of IF systems which are content-based filtering and social filtering. The heritage of the former goes back to Luhn (1958) who introduced the idea of "Business Intelligence System" which identifies every aspect of modern information filtering. However, it was called "Selective Dissemination of Information" at that time. Denning (1982) coined the term "information filtering" in the ACM President's Letter that appeared in the Communication of the ACM. He broadens the discussion which traditionally focused on the generation of information to include reception of information and describes the need to filter emails to separate urgent messages from routine ones. The most influential paper in the 1980s is by Malone et al. (1987) who discuss the two paradigms of IF. One is "cognitive" and the other is "social". The former is equivalent to content-based filtering as discussed by Denning (1982) whereby documents are represented by terms included in those documents.

The most important contribution of Malone et al. (1987) is the introduction of the latter approach (now called "collaborative" filtering) whereby documents are represented by

annotations (e.g. rating) made by other users. It is easy to extract useful features to represent text but more difficult for other multimedia contents such as images, audios and videos. Unlike content-based filtering, collaborative filtering can be applied to documents even if their content cannot be represented in a way that is useful for detection. Similar users can be identified according to their annotation. Documents can then be filtered according to the annotation of similar users.

Some IF techniques are inherited from IR and entail exact match and Boolean model, the vector space model and the probabilistic model. This is no surprise since IF can be considered as a "zero query" approach to IR. Oard (1997) identifies six machine learning techniques for IF systems as rule induction, instance-based learning, statistical classification, regression, neural networks and genetic algorithms. Instance-based learning is of particular interest in this thesis as it has been widely adopted by recommendation algorithms. Instance-based learning represents a family of learning algorithms that compare new documents to documents in training sets to deduce whether they are relevant. Examples of instance-based learning algorithms are the nearest neighbor algorithm, kernel machines and RBF networks (Mitchell, 1997). To classify new documents, they are compared with documents in the training set. The relevance of the new documents is then assigned according to the relevancy of their nearest (most similar) documents (Pazzani and Billsus, 2007).

Recommendation algorithms

Similar to the two key paradigms for IF there are also two key kinds of recommender algorithms as collaborative filtering (CF) and content-based (CB). The first use of the term "collaborative filtering" is in a paper on Tapestry, a mail filtering system (Goldberg et al., 1992). Other recommender systems of that era include the GroupLens Usenet article recommender system (Resnick et al., 1994), Ringo music recommender system (Shardanand and Maes, 1995) and the BellCore video recommender system (Hill et al., 1995). CF is

perhaps the most popular kind of recommender system because it only requires annotation (e.g. rating) of users for operation and is relatively cheap to build. Many successful online recommender systems also rely on this technique. For example, Linden et al. (2003) report on the use of the item-item CF recommender system at Amazon.com.

There are two kinds of CF recommender systems. The first is the user-user CF recommender system. This represents the earliest automated CF recommender system with the tenet to "recommend items which similar users like." All earlier recommender systems such as GroupLens Usenet article, Ringo music and BellCore video utilize this particular type of recommendation algorithm. The underpinning idea is simple. First, users who are similar to the target user must be identified according to the similarity of their past ratings. These users are peers to the target user and included in the neighborhood. Then, ratings of peers are applied to predict the missing ratings of the target user. This method entails computing the degree of similarity between users. A plethora of similarity or distance functions are available and one is selected and applied to the rating matrix. Examples are Pearson correlation, Spearman rank correlation and Cosine similarity. Nonetheless, research has shown that user-user CF recommender systems which utilize Pearson correlation outperform others within this class of recommendation algorithm (Herlocker et al., 1999). More precisely, once peers of the target user are identified, the missing ratings of the target user can be computed according to the average value of ratings made by his/her peers weighted by each degree of similarity with the target user.

To implement a user-user CF recommender system, one must also specify the minimum threshold of degree of similarity of users to be included as peers to the target user or limit the size of neighborhood of the target user. Many authors have discussed this issue. On the one hand, if the threshold of user similarity is too high, then the size of the neighborhood will be too small. This implies that it would be impossible to make rating predictions for many items.

On the other hand, if the threshold is too low, neighborhood size will be larger and many users with a low degree of user similarity will be included as peers to the target user making the rating predictions inaccurate (Herlocker et al., 1999; Anand and Mobasher, 2005).

The second is the item-item CF recommender system. The user-user CF recommender system has the important shortcoming of not being scalable as the user base grows. Similarity between users has to be continuously recalculated for the user-user CF recommender system to work, and this becomes impossible with millions of users. To deploy the collaborative filtering recommender system on large e-commerce websites an alternative algorithm is needed as an item-item CF recommender system. Due to its scalability, this has become one of the most deployed CF recommender systems in use today and Amazon.com also utilizes this type of recommender system (Linden et al., 2003).

Unlike the user-user CF recommender system, the tenet of the item-item CF recommender system is to "recommend similar items to those that users already like", whereby similarity between items is constructed from a rating matrix. The item-item CF recommender system is first discussed by Sarwar et al. (2001) and Karypis (2001). Rather than looking at the similarity between users, this recommender system looks at similarities between the rating patterns of items. In other words, if two items are being liked or disliked by the same users then they are constructed as being similar to one another. Then, target users are expected to have similar preferences for similar items. Again, there are a plethora of similarity and distance functions to select from to construct similarity between items. Cosine similarity appears to be the most popular similarity measure for this recommendation technique as it is simple and can produce good predictive accuracy (Erkstrand et al., 2010). The neighbor of each item (not user) need to be identified and the prediction of missing ratings of items of the target user takes the form of average ratings given by the target user for items in the neighborhood, weighted by the similarity between the items with missing ratings and each

individual item in the neighborhood which the target user has rated. Sarwar et al. (2001) determine the size of the neighborhood at 30 to produce good results on the MovieLens data set.

But why is the item-item CF recommender system more scalable than the user-user CF recommender system? The reason is that the item similarity matrix (as the input into the item-item CF recommender system) lends itself easily to pre-computation, unlike the user similarity matrix (as the input into the user-user CF recommender system). Because users generally rate only a small number of items, the overlap of ratings between one user and another is likely to be relatively small. Therefore, the user similarity matrix can be unstable as the user-user CF recommender system is being continuously updated with new streams of rating data. This is unlike the rating for each item. Overall, each item is likely to achieve many ratings, hence, the rating of one item may overlap more with the rating of other items. Therefore, the item similarity matrix is relatively stable even when the item-item CF recommender system is being continuously updated with new streams of rating data. Because the item-similarity matrix is relatively stable it can be pre-computed and reused when different recommendations are being assembled. This is unlike the user-similarity matrix which cannot be pre-computed and must be recalculated when a new recommendation is made to ensure prediction accuracy; it is unstable and is likely to keep on changing.

The second key kind of recommender algorithm is content-based (CB). Similar to the item-item CF recommender system the tenet of the CB recommender system is to "recommend similar items to those that users already like" (Erkstrand et al., 2010). However, similarity between items is no longer constructed based on a rating matrix as in the case of the item-item CF recommender system. For the CB recommender system, similarity between items is constructed more intuitively based on product features themselves. Recommendation is then made by matching target user preference with product features. For example, if the target user

read and liked fantasy novels before, then Harry Potter ought to be recommended to him/her. Nonetheless, a rating matrix is still an important input into the CB recommender system because the system needs to know the preferences of its target user before it can make predictions and it is this preference data which is stored in the rating matrix. Preferences can be elicited either explicitly by asking target users to rate products or product attributes of items (e.g. genres) that they prefer or implicitly by observing their behaviors (e.g. buying and browsing). A virtue of the CB recommender system is that it does not operate on community rating data; therefore, it does not rely on a large user community for its operation (unlike the CF recommender system). Recommendations can be assembled even when there is only one user if the system already knows his/her preference.

The CB recommender system was originally developed to recommend text documents such as newsgroup messages, news articles and web pages as a well-established technique to automatically extract product features from text documents that already existed. Therefore, it was not expensive to construct and maintain a database of product features for the CB recommender system to operate. Typically, such techniques extract keywords from text documents and these keywords are then used as product features and matched with keywords of user preferences to make text document recommendations.

The standard approach is to automatically transform text documents into lists of keywords that appear within the document. A naïve approach is to set up a list of all the words that appear in the document and describe each document by a Boolean vector, whereby 1 indicates that a word appears in the document while 0 indicates that the word does not appear. One can then try to match a target user profile described by a similar list to the list that describes the document and see if they coincide. If they match, then the document ought to be recommended to him/her. There are two problems with this approach. First, it assumes that every word has the same importance in describing text documents and second, a large

overlap of user profile and item profile will naturally occur with longer documents. Therefore, this approach has a bias for long documents (Jannach et al., 2011).

Thus, a more sophisticated technique is required to automatically extract keywords from text documents. One popular approach to automatically describe documents is TF-IDF encoding format (Salton et al., 1975) which is a product of two terms: (1) term frequency (TF) and (2) inverse document frequency (IDF). Term frequency describes how often certain terms appear within a document. To prevent term frequency becoming higher in longer documents, some normalization should be applied. A relatively simple approach relates term occurrences to the maximum frequency of other keywords in the document (Adomavicius and Tuzhilin, 2005). Inverse document frequency reduces the weight of keywords that appear very often in all documents. These words are not helpful for discriminating documents and more weight should be applied to words that appear in only a few documents. Basically, inverse document frequency is directly proportional to the number of all recommendable documents divided by the number of documents in which certain keywords appear.

Further refinement can be made to the TF-IDF encoding format. First, stop words can be removed. These include articles and prepositions such as “a”, “the” or “on” which appear in nearly all documents. Second, variants of the same word can be replaced by their common stem (root word). For example, “went” can be replaced by “go”. However, there are some pitfalls in this approach. For example, the root word of both “university” and “universal” with completely different meanings is the same as “universe” (Chakrabarti, 2002). Third, phrases can be encoded as words such as “United Nations”. Detection of phrases can be done by looking up manually defined lists or by applying statistical analysis techniques (Chakrabarti, *ibid.*). Nonetheless, there are still some limitations of the TF-IDF encoding format. For example, it does not take into account context (Pazzani and Billsus, 2007). A free text description of a steakhouse might state that “there is nothing on the menu that a vegetarian

would like". In this case, TF-IDF might give high weighting to the word "vegetarian". Therefore, the CB recommender system might mismatch a vegetarian with this restaurant.

Once TF-IDF is computed, similarity can be constructed between text documents. The most common approach is to apply cosine similarity to evaluate whether two documents are similar to one another (Jannach et al., 2011). The neighborhood of an unseen document to be recommended which the user has rated must then be selected. If documents within the neighborhood achieve good rating from the target user, then the unseen document can be recommended to him/her. Again, several variations are possible such as varying the size of the neighborhood, setting a minimum similarity threshold and weighting ratings of the target user by the degree of similarity to compute missing ratings (Allan et al., 1998).

Different kinds of recommender systems have dissimilar weaknesses. On the one hand, there are reasons to believe that the item-item CF recommender system may potentially suffer from popularity bias and become prone to recommending popular items to users. McPhee (1963) looks at two groups of consumers as light consumers and heavy consumers. Consumption baskets of light consumers are monopolized by popular products while consumption baskets of heavy consumers consist of a mixture of both niche and popular products. No consumers buy only niche products. Elberse (2008) shows that McPhee's postulation back in the 1960s still holds in the digital environment. This structure of demand implies that popular products will be assigned as being similar to most products. Therefore, they are likely to be recommended more frequently. Celma and Cano (2008) empirically demonstrate popularity bias in the case of the CF recommender system of Last.fm, a social media platform for music discovery

On the other hand, the CB recommender system suffers from shallow text analysis. First, it may not be enough to look at textual contents alone to make recommendations of, say, web

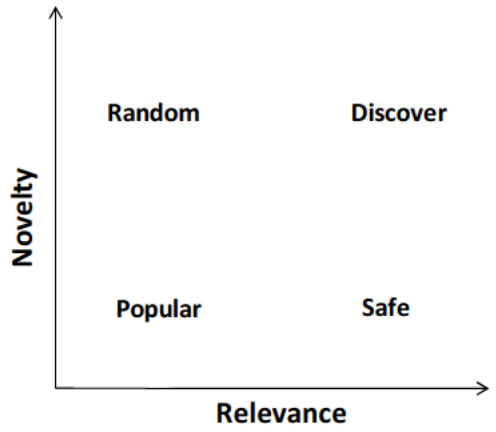
pages. Other aspects such as aesthetics, usability, timeliness and correctness of hyperlinks are also important; however, these are not taken into account by the CB recommender system (Balabanovic and Shoham, 1997). Also, the CB recommender system cannot distinguish between well written and poorly written documents that have the same set of keywords (Shardanand and Maes, 1995) and it is difficult to deploy the CB recommender system in domains of short text documents. Text documents to be recommended may not be long enough to allow a good set of discriminating features to be automatically extracted. A typical example is the recommendation of jokes (Pazzani and Billsus, 2007) whereby it is nearly impossible to distinguish good jokes from bad ones.

Further, both the CF and CB recommender systems suffer from cold-start problems. There are three types of cold-start problem (Erkstrand et al., 2010). First is the item cold-start problem when new items have just been added to the database and have not received enough ratings to be confidently recommended. Second is the user cold-start problem when new users begin to use the recommender system. They have not made many ratings and so the recommender system cannot recognize their preferences. Third is the bootstrap problem when the rating matrix is empty as an extreme intersection between item and user cold-start problems. However, the problem is less serious for CB recommender systems because they rely on product features to construct similarity between items. Therefore, the CB recommender system does not have item cold-start problems. Unlike the CF recommender system, large user communities are not a requirement for the CB recommender system to operate and this requires only one user if the system can recognize the preferences of that user.

There is a belief that creating a hybrid recommender system combining CF and CB will solve these problems because different recommendation algorithms suffer from different shortcomings. However, it is impossible to alleviate one major shortcoming as the inability of

recommendation algorithms to recommend novel items as both CF and CB recommender systems suffer from this. As already discussed, the main tenet of the (item-item) CF recommender system and the CB recommender system is to "recommend similar items to those that users already like". As recommendations from the CF and CB recommender systems are based on selecting products similar to those that users already like, it can hardly be novel. Both systems will generally make safe recommendations whereby relevance is high but novelty is low. However, discovery is about high novelty and relevance as demonstrated in Figure 5.

Figure 5: Trade-off between novelty and relevance



Source: Celma and Lamere (2011)

List of recommended items assembled by recommender systems provides an important mean at which users navigate digital information environment. However, this is not the only way at which digital information is being organized in. Limitless numbers of metrics can easily be assembled from behavioral data and these can be used to organize information as well. To this end, information organization theory as proposed by Weinberger (2008) is useful as it provides a broader perspective to digital information organization. The following section review information organization theory as proposed by Weinberger (2008)

Information organization theory of Weinberger (2008)

Weinberger (2008) discusses the history of information organization and proposes three orders. The first and the second orders are bounded to the physical world; the former orders the things themselves and the latter covers paper-based tools for information organization. The third order is about information organization in the digital realm which is apparently overwhelmed with disorderliness allowing for limitless ways of information organization. In other words, information organization in the digital realm is interactive, with more flexibility compared to the first two orders.

The first and the second order involve organizing things in physical words. The first order concerns organizing the things themselves where a physical thing can only occupy a certain place. For example, a book can only occupy a space on a book shelf and cannot be in multiple places (unless there are multiple copies of the book) as physical nature prohibits that. The second order is about organizing first-order objects with paper-based tools. A prime example is a card catalog which may contain information about books in a library or artifacts stored in a museum. This card catalog contains information regarding the physical location of the first-order objects along with other metadata such as names of the first-order objects and related descriptions.

Information organization in the second order is more flexible than in the first. Card catalogs prepared for libraries may contain metadata concerning books such as call number, authors' names, titles, publishers, place of publication, date of publication, number of pages, size, ISBN and subject heading. This allows the books to be sorted in different ways. Thus, the flexibility of information organization in the second order is more likely to satisfy the needs of visitors to the libraries. However, the flexibility of the second order is still very limited as not much information can be written on card catalogs. Typically, visitors to libraries do not find information regarding how well books sold, whether they are banned in any countries

and their ratings and reviews. Although information organization in the second order contains more metadata than the first, it is still limited.

In the physical realm, things can only be in one place and information about first-order objects can only be sorted in a few ways. The strategy to order information in the physical world is to go through each new arrival of first-order objects one by one and put them away neatly. As new arrivals come in they are instantaneously classified. This is called filtering and crucial decisions must be made straight away about what information to retain. Because the ways of organizing information are limited in the first two orders, only experts with years of educational and professional experience are entrusted to classify first-order objects. Alas, experts always have a bias; they need to work their way up the social institutions they work in and may succumb to influences from, say, funders of their institutions. In other words, they mold their classifications to serve their funders and this may not be suitable for other stakeholders. Indeed, organizing things in one way always disorders them in another in the first two orders.

The third order is about information organization in the digital realm. The key strategy of organizing information in the third order is to include and postpone. All information concerning first-order objects is incorporated but classification is postponed until information is required by the user. Digital technology makes this possible as seemingly limitless information can be attached to an object and the size of the catalog is no longer a problem. With more information, disorderliness increases and things can be sorted and classified in seemingly unlimited ways because of the vast amount of attached metadata. Information organization in the second order has some flexibility but this greatly increases in the third order. Thus, information organization of the third order is better suited to satisfy the needs of multiple stakeholders.

There are four strategic principles for organizing information in the third order. The first is to filter on the way out, not the way in. There is attached value to the abundance of metadata concerning each object in the digital realm. Filtering on the way to decreases the value of this abundance and may rule out information which might be of interest to some people. The second is to assign objects as members of as many categories as possible. People categorize things in different ways and this helps information organization of the third order to better serve the needs of multiple stakeholders. The third is that everything can be treated as metadata. For example, every word in the books can be treated as metadata for those books. The fourth is to give up control. With the abundance of metadata attached to each object, a myriad of relationships may emerge out of the apparent disorderliness and no one person can organize the information in countless useful ways. Thus, it is better to let users take charge by, for example, tagging objects and creating their own bookmarks or playlists. As this social data created by users are shared, this further increases the abundance of metadata of the third order, and, so, expands the possibility of organizing information. Experts are now no longer in charge of information organization in the digital realm. The amount of information is so overwhelming that any expert system will crash. There are just not enough experts to classify everything being generated in the digital ecosystem. Previously, it was the duty of experts to filter information for users to absorb passively. Now, however, users need to be more active and filter information themselves in the third order.

The third order of Weinberger (2008) provides a wider perspective on information organization than IF/IR and recommender system literature. Although recommendation algorithms assemble items for each user differently, there are many more ways in which items can be organized in the third order depending on how much metadata is being attached to those items. For instance, as for books, these metadata may include the book's call number, authors' names, titles, publishers, place of publication, date of publication, number

of pages, size, ISBN and subject heading as well as sales data, whether the books are banned in any countries, ratings and reviews they receive and more innovative ways to organize information on the Internet such as tagging. Users can then interactively explore books along those different verticals. As the amount of metadata attached to books becomes seemingly limitless, the ways at which books can be organized also becomes infinite. This means that the digital information environment can serve different stakeholders over time as their needs change. If these stakeholders want safe recommendations, they can rely on the recommender system, but if they want to make novel discoveries they might need to unravel the full power of the third order and organize information to navigate items in seemingly limitless ways.

Interface

Social media apply algorithms to their databases of social data (predominantly behavioral data) to construct the interface for user interaction. In other words, social media construct and organize their information displays by combining algorithms with predominantly behavioral data. As discussed above, presenting users with ranking is superior to a set of items; ordered lists emerge as the predominant way in which information is organized by social media. Behavioral data input into social media algorithms is 'discrete and granular'. Basic computational procedures such as counting produce numbers can be used to construct an ordered list. Construction of similarity between items to be recommended also results in ordered lists of similar items. Finally, ordered lists are also outputs from recommendation algorithms with items ranked according to the predicted preferences of users.

These ordered lists have real implications on user behavior. Butler et al. (2014) point out that lower participation costs and higher topic consistency cues can increase community size and resilience. Social data can be used to reduce participation cost and heighten topic consistency cues in numerous ways. For example, Butler et al. (*ibid.*) note that allowing messages to be sorted by the number of replies or recency decreases participation cost while revealing the

number of replies also increases topic consistency cues as this helps users to gauge which topics the community finds interesting. Ren et al. (2012) performed experiments on users of MovieLens, a web-based movie recommendation site where members rate movies, write movie reviews and receive movie recommendations. They point out that an increase in identity-based attachment to a group within an online community or bond-based attachment to an individual member of the community increased attachment to the community as a whole which, in turn, increased member participation and retention. Users of MovieLens are assigned into different groups. Information is derived from social data such as top movies rated by different groups, top movies rated by different groups with low ratings from other groups and numbers of new ratings in the past week by each group. These are intended to foster identity-based attachment and competition between groups. Rating agreement and disagreement between users are also provided to foster bond-based attachment.

Ghose et al. (2012) point out to existing literature which demonstrates the benefit of achieving high ranking. Most people start browsing from the top of lists, so higher ranked items are likely to receive more attention. This effect has been documented in various contexts such as food, beverage and elections. In the online environment, it has also been demonstrated that links with higher ranking are more likely to be clicked by users. This ranking effect exists because effort is required to scroll down lists of items. This can be interpreted as search cost. Ghose et al. (*ibid.*) interestingly point out that this search cost is particularly high for mobile phones, so the ranking effect becomes stronger through mobile devices. They determined the negative and statistically significant relationship between the rank of a post and clicks of that post to be much stronger for mobile users than for PC users. A small screen increases the ranking effect as it acts as a serious obstacle to navigation. This finding is particularly important for social media users as a significant proportion access

social media through mobile devices. For example, Facebook currently has 2 billion active users and around 1.7 billion access the site through mobile devices.

While the effect of output recommender systems has yet to be examined in the information systems and media literature, some have already investigated the effect of using a similarity network (or ranking) as a by-product of recommender systems to organize online information. The oldest example of a similarity network is the co-purchase network of Amazon ('Customers who bought this item also bought …') which makes product complementary relationships explicitly visible. Oestreicher-Singer and Sundararajan (2012a) discover that categories whose books are more highly and evenly influenced by a similarity network have consistently flatter demand and revenue. On the other hand, Oestreicher-Singer and Sundararajan (2012b) attempt to empirically estimate the implication of this visibility on demand correlation. They find that visibility increases the demand for products and their complementary products by as much as threefold. Demand correlation exists even without a co-purchase network being made explicit because the goods can be complementary to one another. However, with a visible co-purchase network, the strength of demand correlation increases by as much as threefold.

Social media often have social networking functionality as users can browse through activities which their friends on online social networks perform and this can cast implications upon their behaviors. According to Haythornthwaste (2002), latent ties are created across each and individual pairs of users in a social network maintained by social media. The potential is huge given the size of popular social media sites such as Facebook. This latent tie can potentially be converted into a weak tie, which, in turn, can potentially be turned into a strong tie. Online social network, once created, shapes the kind of information users consume. Wohn and Howe (2016) assess the importance of online social network member diversity (age, race, nationality, occupation, etc.). Users who do not have a particular kind of

diversity in their online social network are likely to be unaware of issues related to that diversity. For example, people with no ethnic diversity are more likely to be unaware of certain ethnic issues.

Anderson (2006) is optimistic about the benefit of similarity networks that can increase demand for niche products. However, not everybody agreed. Goldenberg et al. (2012) point out that the problem of similarity networks is that they are often constructed based on some kind of affinity. Because aggregated data from users (social data) is often used as input for construction of similarity networks, it tends to be successful at connecting products perceived as being similar by users, rendering the network less useful. They analyze the similarity network of YouTube and point out that 56% of the videos connected by the same similarity network are in the same category. On the other hand, Celma (2010) evaluates the similarity network constructed by CF recommender system of Last.fm and suggests that it suffers from popularity bias. Plays counts of different Last.fm artists are strongly correlated with play counts of similar artists. Further, popular artists are more likely to act as hubs within a similarity network linked to many other web pages.

Fortunately, the integration of a similarity network with a social network into a dual network can alleviate this problem. People create links to a similarity network as they participate within the online community (comments, reviews, posts, etc.). These links can be seen as their personal recommendations regarding each product, which may complement the similarity network and help platform users to better discover contents. Goldenberg et al. (2012) also investigate this using data from YouTube. They find that profile pages of users have unique structural properties, making them better content brokers than similarity networks of YouTube, as users are likely to post links to more varied contents, bridging different product circles. For YouTube, less than 20% of users generated links connecting products of the same category and these structural properties cannot be easily replicated

algorithmically by comparing the actual dual network of YouTube with the synthetic version. They also determined that a dual network enables users to find satisfactory contents more quickly and replicated their YouTube study with data from Last.fm with similar results.

Overall, Aristotelian classification is in decline and prototype classification is on the rise as information organization shifts online. Weinberger (2008) suggests that things can only be organized in a few ways in the physical world. The essential idea is that things have a clear-cut boundary. However, things can be assigned to be members of many categories and they can be sorted in seemingly limitless ways in the digital environment by prototype classification where boundaries of categories appear fuzzy which is the key principle to classification in the digital ecosystem.

Digital objects

Smith (1996) notes that in the late 20th century the current theories of computing failed to do justice to digital objects which increasingly populate everyday life including blogs, wiki, web pages and personal profile pages. The conceptual apparatus of analytic science and philosophy remained inadequate to deal with these objects as only a pretheoretical understanding existed. People knew about blogs, wikis and web pages but no theories of digital objects were postulated, and there was no unitary answer as to the status of digital objects.

Fortunately, there is now a small but growing literature on theories of digital objects (Kallinikos et al., 2013). This section discusses the concepts and is divided into three parts. First, the numerous papers on the theory of digital objects are discussed followed by the theory of digital objects as presented by Kallinikos et al. (2013) which synthesizes the pre-existing literature. Third, the theory of digital objects is applied to elucidate the third order of

Weinberger (2008) and connect the theory of digital objects with the theory of information organization and the theory of classification, the other two theories relevant to this thesis.

Research papers on digital objects

Ekbia (2009) studies bug fixing in Free/Open Source Software development using a processual perspective whereby a bug is seen as a manifestation of a digital object. He develops a theory of digital objects by analyzing the network and process that they trace and mediate. Digital objects are constructed through collective activities of justification whereby they are constantly being discussed, contested and negotiated. Ekbia concluded that digital objects are better seen as quasi-objects as they lack stability. Although qualification is a central activity in all situations of uncertainty, discord and disagreement, digital objects are subjected to a different kind of qualification from other familiar objects of daily lives. Likewise, Kallinikos et al. (2010), Kallinikos and Mariategui (2011) and Manovich (2001) study digital objects such as files, images and videos often embedded within complex, distributed and shifting digital ecosystems and pointed out that they are fluid and editable.

From an economic perspective, Faulkner and Runde (2011) look into non-material technological objects which do not have a physical mode of being. Examples are numerous information goods (Shapiro and Varian, 1999) such as product designs, sales reports, mathematical algorithms and computer files. They further suggest that it is important to distinguish non-material technological objects from their bearers in which non-material technological objects are inscribed. There are material bearers (e.g. books, newspapers, computer printout) and non-material bearers such as bitstrings which have at least three properties. The first is non-rivalry, as uses by one person do not affect simultaneous uses by others. The second is seemingly infinite expandability as they can be made available to other users at very low marginal cost and the third is a high degree of recombinability which encourages activities like mash-up and reuse of codes.

Yoo et al. (2010) evaluate the loose coupling between the device layer, network layer, service layer (application functionality) and contents layer of digital objects. Digital objects are reprogrammable which leads to a separation between the device and service layers and entails homogenization of data whereby digital contents (audio, video, text, image etc.) are digitalized leading to a separation between network and content layers. Layered modular architecture has emerged and this has unleashed recombinant innovation as a distinct form of digital innovation. Here, digital innovation happens when digital and physical components of digital objects are combined to produce novel products. Unlike traditional modular architecture, components from different design hierarchies can be brought together to create a kind of digital object (not merely a difference in degree), unleashing generativity whereby innovation can spring up independently at any layer leading to cascading effects to other layers.

According to Benkler (2006), Lessig (2006), Zittrain (2008) and Kallinikos et al. (2013) digital objects can be characterized along similar lines to digital infrastructure as essentially end-to-end architecture since they feature modularity as well as granularity. Zittrain (2008) proposes that the Internet is constructed according to the Procrastination Principle with simplicity and intentionally left incomplete as it is thought that most problems confronting networks can be solved later. An end-to-end argument is made that most features in a network ought to be implemented at computer endpoints rather than in the middle by users themselves as problems or needs arise. Modularity and loose-coupling between components which constitute the network enable a clear division of labor among developers who work to improve the overall system. There is no need for explicit coordination between them as they can make their contribution without knowing what others are doing or how exactly other components operate.

A theory of digital objects

Kallinikos et al. (2013) proposes that all digital objects share four generic attributes making them less stable and more malleable compared to non-digital objects such as physical entities and cultural records (e.g. paper-based tools). First is editability as they can be continuously modified. Digital objects are composed of separable and clearly differentiable elements and they can be modified in various ways. For example, new elements can be added, existing elements can be deleted and the functionalities of elements can be altered. This attribute is related to databases which are often required to be regularly updated. Second is interactivity which allows for contingent exploration and alternative pathways to be activated by users. An example to illustrate this attribute is a dynamic website. This attribute also makes digital objects malleable and less stable; it is distinct from editability in that it is not about modification of digital objects themselves. Third is openness as digital objects can be reprogrammed by other digital objects and not only those that govern behaviors. An example to illustrate this attribute is modification of digital images by picture-editing software. Again, this is different from editability as it requires external interferences. Last is distributedness as digital objects are often merely temporary assemblies of more basic elements and often distributed across the whole information infrastructure. Elements are seldom contained within a single location and they lack clearly identifiable borders. This heightens the importance of assembly procedures in relation to standalone elements. New digital objects can be constructed simply by innovatively combining existing elements in new ways.

These four attributes presuppose more fundamental constructs. The first is modularity. Ever since Simon (1969), modularity has been associated with systems that have a loosely-coupled network of functional relationships between numerous self-sufficient blocks mediated through interfaces. Modularity has been associated with the realization that integral, *en bloc* objects or systems are hard to act upon, control and manipulate. While modularity is relevant

to both physical and digital objects, it runs much deeper and wider for digital objects and technologies. According to Yoo et al. (2010), physical objects obey fixed design principles whereby interfaces of their components are functional specific and constitute single product design hierarchies. For example, spare parts of a vehicle can seldom fit into other models. In contrast, digital interfaces can accommodate much wider spectrum functions and are often designed to be function agnostic. Second, more importantly for this thesis is granularity. Although modules are seemingly *en bloc* they can be further decomposed into fundamental elementary units of the binary. Each of these binaries is separable and clearly differentiable from one another in contrast to analogue systems and allows each to be independently manipulated to bring digital objects into new configurations. There are two operations which set granularity distinct from modularity. First, with granularity digital objects can be traced down layer by layer because of their binary status. This enables a database to be data mined and pictures to be edited by image editing software. Second, granularity enables piecemeal intervention such as the widespread practice of digital content editing in Wikipedia. The piecemeal nature of digital objects enables people to contribute as a collective pursuit according to their time availability, capacity and inclination.

Information organization, social media & the theory of digital objects

Interface, information display or information organization in the digital realm is a digital object constructed by two digital components as a database and an algorithm which operates on the database. A database within the digital ecosystem is constructed on the editability and distributedness of the digital object. Databases collect and store all metadata related to items to be organized (editability). According to Weinberger (2008), everything connected to items being organized can be treated as metadata. This can be behavioral data such as consumption data and tagging data as discussed by Alaimo and Kallinikos (2016) as well as metadata derived from contents themselves such as words in books and characteristics of music e.g.

beats and rhythm. A database needs to be constantly updated as new metadata is being created (editability). For example, behavioral data is being created all the time with each user interaction. This newly created behavioral data has to be collected and stored in the database. Further, the database may source its metadata from all over the digital ecosystem, demonstrating the distributedness nature of digital objects. Aggregation of this metadata is not an easy task with no centralized authority to produce and maintain handholds for each item being organized. This needs to be managed to effectively organize information in the digital realm.

The amount of metadata attached to each item to be organized in the digital realm is much greater than metadata attached in the physical realm on, say, card catalogs. The amount of metadata resembles disorderliness which powers information organization of social media. Prototype classification is used in the digital ecosystem whereby an item can be placed in as many locations as possible. Items in the database can be sorted in seemingly limitless ways according to the metadata with the help of an algorithm and prototype classification emerges from this. This renders information organization in the digital realm highly interactive as users can always choose an alternative path for exploration. For example, items presented to users can be personalized according to their preferences or they can simply be sorted by popularity. Lastly, the algorithm is characterized by openness and may be occasionally altered to produce a better personalization algorithm. Overall, the characteristics of digital objects render information organization in social media unstable and in constant flux. Databases are constantly updated and sometimes the algorithm is altered.

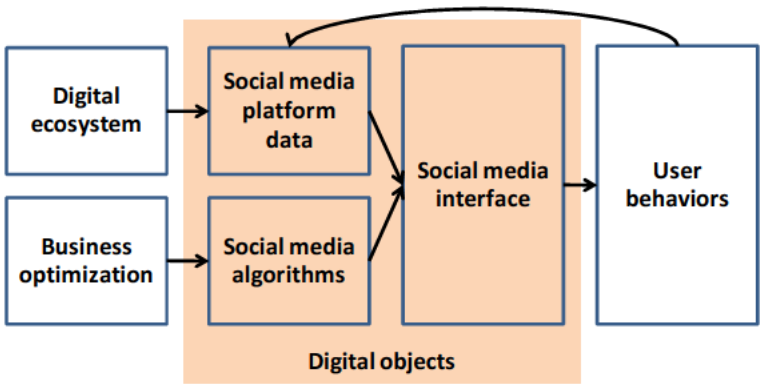
Theoretical synthesis and conjecture

Figure 6 depicts the inner operational aspects of social media which collect social data from users. Simultaneously, social media sites are transformed into social media platforms; they also source their data from the broader digital ecosystem. The collected data has to be

aggregated to be useful. This cycle demonstrates the complexity or the difficulty to distinguish the operational aspects or the production side from the consumption side of social media because once user behaviors (use/consumption) are shaped by the social media interface and captured as social data they are fed back into the inner operation of the social media platforms. Nonetheless, such distinction is useful analytically as operational aspects have been identified as a fruitful area of research.

Social media algorithms then operate on social media data to construct a social media interface. Social media sites have the status of perpetual data as they always strive to maximize user engagement and business values by altering their algorithms. Social data can be used to construct social media interfaces, information displays or information organization to assemble seemingly limitless numbers of ordered lists. Users can then interactively browse through these ordered lists and prototype classification automatically emerges. Because social networking functionality is often embedded into social media, users can also browse through profiles of their online friends to discover things of interest. In turn, user behaviors are shaped into social media interfaces which render them unstable because social media interfaces and the components which construct them (database and algorithm) are essentially in flux. While databases are constantly updated, algorithms can change as social media optimizes their operation, rendering social media interfaces also unstable.

Figure 6: Theoretical framework



Numerous conjectures result from this theoretical framework which are evaluated by papers in this thesis either by simple descriptive statistics or hypothesis testing. First, *navigation of information display as assembled by social media is highly interactive*. There are infinite ways to organize social media information through sorting and retrieval. Second, *information organization of social media is highly unstable which would also render user behaviors unstable*. This occurs because information display and components which underlie its construction are digital objects which are in constant flux. Third, *quality of data aggregation has significant implications on user behaviors*. Data aggregation is very important, especially for social media platforms which source data from the wider digital ecosystem. This is because of distributed labeling whereby an object can be termed differently by diverse users. Fourth, *the amount of data captured by social media platforms limits the usefulness of their information displays*. Basically, the more data captured the better and more useful is the information displayed and assembled by social media. This is because there are always more items to be put on display and functionalities such as recommendation algorithms can only work with large amounts of data. Fifth, *output from the recommendation algorithm (recommendation list) has real implications on user behaviors*. Generally, recommender systems "recommend similar items to those that users already like". Although this is only one of the ways in which information can be organized it is often featured prominently by social media sites and has significant implications on user behaviors. Sixth, *circles of friends on a social network can influence the behaviors of users*. Social media is often embedded with social networking functionality which allows users to browse through profiles from their friends which influences their behaviors. Seventh, *metadata attached to items displayed has influence on the behaviors of users*. There are many kinds of metadata attached to items displayed on social media, some are relevant to users and so cast real implications.

Research methodology

The use of analytical techniques to make sense of data can be traced back to the 18th century. The obvious difference today is that we are living in a data-rich environment with online economic and social transactions captured as digital data. Big data has emerged, causing changes across academic disciplines as answers to new sets of questions become possible. This thesis analyses specifically the social big data created as users participate on social media platforms or, more generally, as any social transactions are being recorded as digital data. Data scraped from Last.fm, the empirical object of this thesis were subjected to quantitative analysis to better understand how social media platforms like Last.fm assembles information for its users.

This section contains four parts. First, the unsettling impact of big data on scholarships across academic disciplines is discussed, followed by screen scraping or the technique which allows researchers to sample and analyze social big data. The methodology employed in this thesis as basically screen scraping and analysis of social data is compared and contrasted with other research methodologies and their pros and cons are evaluated. Finally, research which deploys social big data across various domains is discussed.

Big data and scholarships across academic disciplines

Savage and Burrows (2007) generated considerable debate among sociologists by arguing that empirical sociology is facing a coming crisis with the emergence of social big data. Fifty years ago, sociologists and social scientists occupied the apex of the social research apparatus such as surveys, interviews and ethnography. Now, their position has become insecure with the emergence of social big data. Empirical research methodologies of social scientists no longer allow them to access the 'social' in ways that interested groups find valuable. Sampling is the key to conducting good surveys and interviews as it allows researchers to

make inferences about the population through collection of data on a small number of people (Savage and Burrows, 2009). However, their response rate keeps on falling as people no longer feel honored to be asked for their opinions. Alas, this is not an issue for organizations which capture and wield social big data and allows them to gain access to data of each and every individual in the whole population.

Lazer et al. (2009) urge social scientists to turn to computational methods and advised that the emergence of computational social science has been much slower compared to the way that big data has transformed other fields such as biology and physics. Computational social science is already occurring but it has been mostly restricted to technology companies (e.g. Google and Yahoo) and government agencies (e.g. the National Security Agency). To foster a stronger computational social science community, industry and the academia must collaborate to facilitate research, enforce privacy and provide liability protection for corporations. On the other hand, Conte et al. (2012) suggest many fruitful research areas including modeling the layers of complexity in the real world which include not only micro and macro layers of complexity but also intermediate layers (e.g. groups and tribes). Another suggestion is to model culture. Axelrod (1997) attempts to address the problem of cultural dynamics and states, "if people tend to become more alike in their beliefs, attitudes, and behavior when they interact, why do not all differences eventually disappear."

Digital humanity is another discipline that is being transformed by big data. Originally, digital humanists worked in digitalization and archiving projects to transform cultural objects into digital forms but now they increasingly curate and analyze data of digital origin (Schnapp and Presner, 2009). Manovich (2011) points out that the largest data sets used by digital humanists are still relatively small. Humanists still work on their desktop computers using standard software but data size will grow exponentially once they start to work with born-digital user-generated content. Similar to the case of computational social science, the

challenge for digital humanists is to access data, as only social media companies have exhaustive access to social big data. However, fortunately, there is a way around that as digital humanists can gain partial access to social big data through application programming interfaces (APIs). For example, Manovich (*ibid.*) uses a Flickr API to download 167,000 images from "Art Now" Flickr group for analysis.

Information systems (IS) scholars are particularly well positioned to wield big data for their research as the IS discipline has the longest history of conducting research regarding digital technology and data in society and organizations. For example, the field of management information systems (MIS) has emerged with business processes automated by digital technology and business transactions recorded as digital data. Furthermore, IS scholars understand the complexity of the infrastructure needed to handle big data (Agarwal, and Dhar, 2014). The success of IS scholars is reflected in the fact that their works have been published by widely read scientific outlets such as PNAS and Science (Aral et al., 2009; Aral and Walker, 2012). Agarwal and Dhar (2014) suggest that IS scholars should look beyond traditional journals and communicate with the larger community of scientists and businesses in the age of big data.

Accessing social big data: screen scraping

Google founders explained in their classic article "The PageRank citation ranking: bringing order to the Web" that the web is a vast collection of "completely uncontrolled heterogeneous documents" in terms of both internal variations and external meta information (Page et al., 1998). Screen scrapers emerged as devices capable of bringing order to the web. They turn the heterogeneous mass of online data into formatting information as web data is progressively stripped of its useless elements and formatted to produce a well-ordered, usable data set. Redundant html code and other irrelevant bits of data are removed until only targeted data remain (Marres and Weltevrede, 2012).

Figure 7: Timeline for web scraping

Source: Krijnen, Bot and Lampropoulos (2014)

Figure 7 reveals the timeline for web scraping (Krijnen et al. 2014). Large horizontal bars represent developments that occurred over multiple years without a clear starting point. Named events surrounded by circles represent certain milestones linked to a specific horizontal bar. The rise of screen scraping coincided with the rise of the Internet, driven by the publication of scraping specific libraries but slowed down through rising legal tension. Official APIs emerged as an alternative channel to scrape web contents and open source communities emerged with the Internet. Many scraping libraries have become available over the years with a wide variety of programming languages. Beautiful Soup is a library designed by Python and released in 2004. It is considered a milestone as the most sophisticated and advanced global library for web scraping. Another factor that boosted the deployment of scrapers was the emergence of scraping software with visual interfaces in the early 2010s. Kimono Labs was launched in 2013 as another milestone[1]. Arguably, this development democratised knowhow for the business of screen scraping. Web APIs started to appear in the early 2000s with the first being salesforce.com. Web APIs provide an alternative way of

[1] Kimono Labs was acquired by Palantir in 2016 and ceased to offer publicly available cloud service

gathering data. Compared to scraping html front end where users see, web APIs provide easier and more consistent access to information. However, one of the disadvantages of scraping using APIs is the likelihood of rate limits imposed on developers.

Alas, widespread deployment of screen scraping has been slowed down by legal tension. Web scraping has been subjected to many lawsuits for two major reasons. First, web scraping often involves collecting data which are core raw material for some technology companies. Second, automated web scraping can be taxing on the servers of the targeted websites. The first big court case was in 2000 concerning eBay and Bidder's Edge (BE), an aggregator of auction lists which included a substantial number of auctions that originated from eBay on its site. When technical measures to prevent web scraping failed, eBay sued BE. The court judged in favor of eBay since even though BE's use of the eBay bandwidth was small, this could potentially harm eBay since other companies might follow BE's example and scrape data from eBay. Another major court case was between Facebook and Power.com in 2009. Power.com offered a website that enabled its users to aggregate data about themselves that also spread across various social media sites. Facebook sued Power.com on the accusation that Power.com scraped copyrighted materials from Facebook. Although Facebook did not own the copyright of its users' profile data, it argued that it owned the copyright of the website framework surrounding the users' data and that scrapers of Power.com copied the entire web page from Facebook to extract user data and thus made use of copyrighted material. In this case, the court judged in favor of Facebook.

Marres and Weltevrede (2012) point out that screen scraping provides a number of distinctive affordances for social science. First, screen scraping renders data on the Internet in a format that is usable for social science. Arguably, screen scraping unlocks the 'sociological potential' of the web. Screen scraping has the potential to make very large quantities of user-generated data amassed on social media sites (social big data) available for social scientists.

Second, screen scraping may potentially resolve the long-held research problem raised by online digital data which is often referred to as the problem of 'dirty' data (Bollier, 2010). Web data is often described as 'incomplete', 'messy' and 'tainted' (Savage and Burrows, 2007; Uprichard, 2012). This is especially true when one compares online digital data with other data sets that social scientists use such as survey data and interview transcriptions. Fortunately, screen scraping can be used to extract structured information from heterogeneously formatted data online. It is also possible to assemble random samples of objects (e.g. users) processed by social media sites so that inferences can be made for the whole population of the objects by analyzing a smaller set of samples (Gjoka et al., 2011) as these objects are often assigned with unique identifiers. Thus, one can easily assemble random samples of objects, including individual objects by naming them with random unique identifiers.

Social big data vs. other research methodologies

Research methodologies often distinguish between intensive research strategies that capture the locomotion of social relations 'in process' and extensive strategies that capture the structure of social relations at particular moments and are therefore 'punctiform' in providing a snapshot of these relations (Sayer, 1992). Arguably, social science research is characterized in terms of a trade-off between extensive and punctiform research which captures variation at the level of populations, but only at specific moments and only retrospectively (e.g. national surveys), whereas intensive research captures social processes but only in very specific social contexts or amongst particular social groups (e.g. workplace, ethnography). According to Edwards et al. (2013), the promise of social big data is the ability to employ extensive research strategies to investigate social processes.

The methodological matrix presented in Figure 8 presents social big data (or social media analysis) in terms of extensive/intensive research strategies vs. locomotive/punctiform

research design along with more traditional research toolkits of social scientists such as ethnography, interviewing, surveys and experiments. Social big data processed by social media sites capture the locomotive states of the whole population in real time rendering everyone in the system visible as events unfold. Alas, social scientists working outside social media companies are unlikely to have the privilege to gain access to all the social big data possessed by companies in real time. Nonetheless, it is possible for them to assemble (as already discussed) and analyze a representative sample of users and, in some cases, the actions of users may have timestamps, enabling users to be analyzed through time and allowing for extensive research strategies with locomotive research design implementation.

Figure 8: Methodological matrix

		Research design/data	
		Locomotive	Punctiform
Research Strategy	Intensive	e.g. Ethnography/Participant Observation	e.g. Qualitative interviewing
	Extensive	Social media analysis (capturing naturally occurring data in real time at the level of populations)	e.g. Surveys and experiments

Source: Edwards et al. (2013)

Edwards et al. (2013) further point out that the use of social big data can also re-orientate research questions asked by social scientists. For example, social scientists working with social big data can no longer theorize based on conventional individual attributes such as race, age, class, gender and offline context because social big data often do not contain this information. Digitalization has transformed the society and social scientists now require a different set of theories; maybe social big data can be better suited to answer the questions that emerge from these theories. For example, object orientated sociality has become the new 'digital publics', whereby people convene around particular knowledge or cultural objects such as parenting, job seeking, dating and celebrity gossips. These objects help researchers to

better understand new forms of social organization, change and identity than conventional individual attributes and this information can be readily found in social big data (Knorr-Cetina, 2001). Likewise, with non-pyramidal and non-hierarchical structures inherent in online communications (Spears and Lea, 1992), the nature of inequality may change and the characteristics of digital online elites (e.g. the "Twitterati") may no longer map onto the characteristics of traditional elites.

Social big data provides a great opportunity to research the operational aspects of social media as the very behaviors of social media users captured as social big data. Alaimo and Kallinikos (2016) point out that social media sites do not nurture social interaction neutrally as they are artificially personalized technological environments. How do social media sites assemble these environments and what are the effects on the behaviors of users interacting on these sites?

Research using social big data

Some examples of research that intensively deploy social big data are presented below. They span both information systems and other disciplines (e.g. marketing, crime analysis, epidemic intelligence). The utilization of social big data for information systems research and research in other domains is discussed.

Information systems research

Besides research on the operational aspects of social media (e.g. Anderson, 2006; Celma, 2010; Goldenberg et al., 2012; Oestreicher-Singer and Sundararajan, 2012a; Oestreicher-Singer and Sundararajan, 2012b), social big data analysis has been applied to other types of information system research. The first covers firm studies. Luo et al. (2013) successfully use social media data to predict firm equity value. According to the efficient market hypothesis, new information may change market expectations and change company stock prices

(Samuelson, 1965; Fama, 1970). Social media has become a new source of information which provides timely assessments of a firm's product and brand performance compared to sales information. An association is determined between online consumer ratings and information in blogs regarding equity value. Arguably, ratings and blogs can furnish more relevant product- and brand-specific information than other forms of social media such as videos and networking sites (Tirunillai and Tellis, 2012). On the other hand, Greenwood and Gopal (2015) prove an association between the impact of increased media coverage from two kinds of media on firm founding rates. More specifically, they looked at the three largest blogging platforms (Typepad, Blogger and WordPress) and traditional media coverage from 11 major US newspapers. There are two mechanisms by which media influences behaviors Discourse in media can legitimize topics, industrial sectors and firms (Pollock and Rindova, 2003) and also it creates availability bias, causing decision-makers to systematically over or underestimate the odds associated with events occurring (Sunstein, 2003). Entrepreneurs may respond to increases in discourse by inferring that the entry of a new venture into a widely discussed technology sector will stand an increased chance of survival. Hence, increased media coverage on specific technology sectors will lead to greater observed firm founding in that sector.

Second is the area of user studies. Oestreicher-Singer and Zalmanson (2013) review the related literature and suggested that a ladder of user participation on social media exists which entails four steps as (1) content consumption, (2) content organization, (3) community involvement and (4) community leadership. They further establish that users higher on the ladder of participation on Last.fm, the empirical object of this thesis, tend to subscribe more for paid services. Zeng and Wei (2013) examine the relationship between social connectedness and creation of content on Flickr, analyzing data from 1.8 million users. They calculate similarities between tags applied to photos that pairs of users uploaded over three

time periods as before, around the time of and sometime after a social tie is formed between them. Results show that people tend to upload more similar photos around the time of the formation of a social tie as the formation is driven by similar interests or shared activities. However, thereafter, their photos became gradually less similar and the difference between the popularity levels of users' content moderated this relationship. That said, photos uploaded by similarly popular users diverge much more than those with greatly different popularity levels. The authors use social psychological motivations to explain these results.

Third is the area of employment studies. Lynn (2013) studies the impact of change in employees' network positions before and after the introduction of social networking on worker productivity as measured by billable revenue. Social network theory predicts that an information-rich network that is low in cohesion and spans structural holes is associated with higher work performance. Brokers who bridge these structural holes are endowed with early exposure to novel information and can act as hubs to facilitate information flow between otherwise unconnected groups. Studies have shown that people whose networks are rich in structural holes have a competitive advantage over their peers and tend to receive superior performance ratings and higher compensation (Burt 1992; Lin 2001; Cross and Cummings, 2004; Wu et al. 2009). Lynn further shows that workers can actively manage their network to gain competitive advantages by analyzing the electronic communications of 8,037 employees over two years. The data contained emails, calendar events and instant messages within a global information technology firm.

Other fields of research

Belllo-Orgaz et al. (2016) note that analysis of social big data has been applied to at least three other areas outside information systems which are marketing, crime analysis and epidemic intelligence. Marketing researchers believe that social big data provides the

opportunity for businesses to obtain opinions from vast numbers of customers. Successful cases of deployment of social big data include e-commerce companies like Amazon and eBay. One of the benefits of social big data is that it allows companies to generate more targeted advertising and marketing campaigns. Trattner and Kappe (2013) present results of an ad-driven social network-based marketing campaign centered on Facebook. They demonstrate that ads placed on a user's newsfeed increased the number of visits, profit and return on investment (ROI) of a web-based platform. Many marketing researchers have analyzed Twitter data. Jansen et al. (2009) investigate micro-blogging as a form of electronic word-of-mouth for sharing consumer opinions concerning brands. They retrieve and analyze more than 150,000 micro-blog postings from Twitter and find that 19% mention a brand and almost 20% contain some expression of brand sentiments. They conclude that micro-blog postings can be used to gauge customer sentiments of brands and their competitors in real time. Asur et al. (2010) use data from Twitter to forecast box office revenues for movies. The authors demonstrate that a simple model built from the rate at which tweets are created about particular topics outperform market-based predictors and that sentiments extracted from micro-blog postings can be utilized to improve forecasting accuracy.

Criminals tend to have repetitive behaviors which are dependent on situational factors. The purpose of crime data analysis is to identify these patterns as a means of detecting and preventing crimes. Here, social big data analysis can be very useful in supporting law enforcement agencies. Communication between citizens and government through telephones and face-to-face can be digitalized and analyzed along with already digitalized formats such as emails. Ku and Leroy (2014) propose a decision support system that combined natural language processing techniques, similarity measures and classification approaches to automate and facilitate crime analysis. Filtering and identification of similar crimes can provide useful information to law enforcement agencies to catch suspects and improve crime

prevention. Geographical distribution analysis is also highly relevant to crime analysis. A number of mapping techniques can be used to identify crime hotspots. Chainey et al. (2008) assess these techniques and conclude that kernel density estimation consistently outperforms other techniques. Gerber (2014) exploits spatiotemporally tagged tweets for crime prediction and shows that Twitter data improves crime prediction performance versus standard approaches based on kernel density estimation.

Epidemic intelligence can be defined as the early identification, assessment and verification of public health risks and timely dissemination of alerts (Paquet et al., 2005). This discipline includes automated and continuous analysis of text accumulated on the web, including that pertaining to the realm of social big data. For example, search engine queries were analyzed to track influenza and the relative frequency of certain queries highly correlated with the percentage of physician visits in which a patient presented with influenza-like symptoms. However, this technique to monitor influenza is only applicable to areas with a large population of web search users (Ginsberg et al., 2009). A good exemplar of this is Google Flu Trend which was later discontinued as the estimations appeared inaccurate especially over 2011 to 2013 when flu prevalence was consistently overestimated (Lazer et al., 2014). Nonetheless, later studies continued to find value in Google search queries in estimating flu prevalence (Preis and Moat, 2014). Twitter data is also potentially valuable for epidemic intelligence. For example, Culotta (2010) gathers Twitter messages related to flu and correlates them with the Center for Disease Control and Prevention (CDC) statistics. The author finds that the best model produces a correlation of 0.78 (simple model regression). Aramaki et al. (2011) present a comparative study of various machine learning methods to classify tweets related to influenza into two categories as positive and negative. They show that support vector machine models that used polynomial kernels achieve the highest accuracy and the lowest training time.

Empirical object: Last.fm

Last.fm is one of the oldest and previously most popular online music discovery services. It was established in 2002 and acquired by CBS for $280 million in 2007. Since its inception to early 2009, users can stream free music directly from the service. In April 2009, the company limited free streaming to the US, UK and Germany, citing the inability to recover music licensing fees from advertising[2]; users in other countries were required to pay a subscription fee. Over the last five years, Last.fm has gradually wound down all streaming operations to focus its business exclusively on music discovery. Last.fm operates a 'freemium' business model whereby basic services are provided free and premium services offered for a fee which together with advertising constitutes the main revenue sources of the company.

Last.fm's core activity is to suggest music to its user base by collecting and computing data on user listening behavior taken from partners such as Spotify[3] and YouTube. Last.fm powers its music discovery service with its proprietary technology called 'AudioScrobbler' which is essentially an item-based collaborative recommender system (Ekstrand, 2010; Jannach et al., 2010; Riedl and Smyth, 2011; Konstan and Riedl, 2012) that works by computing data on listening behaviors 'playback events' collected through application programming interfaces (APIs) from more than 600 playback applications, services and devices distributed across the web.

Figure 9 illustrates the role of APIs as boundary resources connecting Last.fm to external devices, applications and platforms and allowing data on user listening behaviors 'artist names' to be ingested in the Last.fm system. The system then 'counts' the 'artist names' data entity and produces 'play counts' (or 'events') (Figure 9 passage 3). 'Play Count' or

2 From Last.fm blog: http://blog.last.fm/2009/04/22/radio-subscriptions

3 From Last.fm blog: http://blog.last.fm/2011/11/30/lastfm-for-spotify

y nt ' 'Is. More precisely,

ha r I s i listening data from

the connected 600 playback applications, a l ınt to point out that

s a ıtity as 'play count'

de La t a s c llaborative filtering

recommendation system.

Figure 9: Last.fm layered architecture

Item- ol bo ve similarity ranking.

In the case of Last.fm, 'AudioScrobbler' personaliz lations by mapping

'playback events' of :ks' on the basis of

i ri l a , Last.fm computes

the probability that users who e s t w d l t s ɔ similar artists.

i m a :s of ordinary social

F s a / user activities as

b nt '; c a us r he rt :ipate in the social

s for a n oi ıg existing groups.

Here, they can start to participate in discussion threads on topics of int :est. Users can also

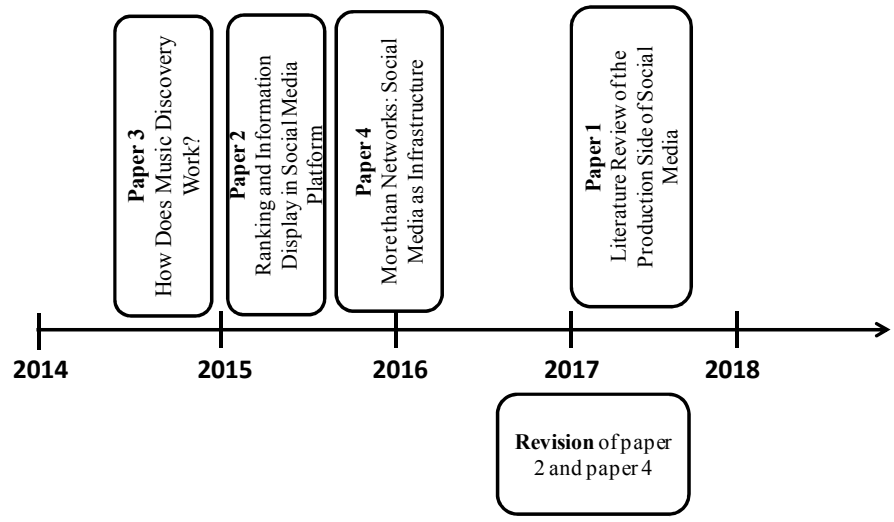
choose to contribute to the platform in more substantial forms by writing blogs about music, writing biographies (wiki) about artists or uploading images, music or videos of their favorite artists. They can also add information regarding music events and invite other users or join upcoming events.

One of the most important activities of users (besides producing listening events) is to 'tag' artists, albums and tracks with any keywords. 'Tags' enter into the 'Audioscrobble' of Last.fm to construct 'similarity networks'. The reason is simple. Computing similarity scores solely on the basis of user listening behavior may be problematic. For example, users who like to listen to classical music may also like to listen to rock music. A score based solely on listening data would determine classical and rock music as similar to one another. To attenuate this problem, artists on Last.fm are deemed similar to one another not only when they are listened to by the same group of users but also when they are labeled with the same 'tag'.

Research procedures and research contributions

In this section, four individual papers are summarized followed by a description of how social media organize information. Figure 10 outlines the temporal interrelationship between the four papers. The third paper was written during the second half of 2015, followed by the second and fourth papers. The first paper was written at the first half of 2017. Concurrently, I also revised the second and the fourth paper during the period.

Figure 10: Temporal interrelationship between the four papers



Individual papers

Paper 1: Literature Review of the Production Side of Social Media

This paper presents a systematic review of the literature on the production side or the operational aspects of social media or its inner workings behind the user interface. It examines eight information system journals (Senior Scholars' Basket of Journals) and ten media journals related to technology between 2000 and 2016. A total of 1,732 papers focused on the use of social media or the consumption side with only 99 papers identified on the operational aspects which comprise a relatively small but growing area. The number of papers published on the consumption side began to gain traction by 2007 and accelerated further in 2013. However, research on the operational aspects has grown steadily since 2013 and presents a fruitful area for further studies. The inner workings are of importance for the operation of social media as they cast real implications on users' behaviors.

Papers on the operational aspects were then furthered classified into four categories as algorithms (16 papers), data (26 papers), interfaces (54 papers) and reaction against the commercial aspects of social media (3 papers). Predictably, there were fewer studies on algorithms than interfaces; the latter are more visible while the former are seemingly invisible

and are often viewed as opaque or 'black boxes' by information systems and media scholars. Although algorithms have been widely studied by computer scientists, they often conduct research in a technical manner and, hence, do not uncover the societal implications of information systems which are deemed important by media scholars. Papers concerning the operational aspects in each category are then reviewed with future research directions discussed.

Paper 2: Ranking and Information Display in Social Media Platforms

This article applies the theory of digital objects as postulated by Kallinikos et al. (2013) to elucidate the information display techniques applied on social media platforms. It posits that social media information displays are assembled using two components as databases and algorithms. Algorithms operate on the social data collected to assemble information displays for user interaction. Thus, in turn, generates even more feedback into the social media database. Three theoretical claims regarding digital information displays have emerged. First, ranking is recognised as the dominant strategy to display information on social media platforms. Second, navigation through the information display is an intensely interactive experience. Third, the information display is dynamic, fluid and unstable. These have culminated as the interaction system discussed by Wegner (1997). The information display strategy of Last.fm, a popular social media platform for music discovery is then presented and analysed to assess these three theoretical claims. The first two theoretical claims are confirmed by data from Last.fm, while the assessment of the last claim is surprising as the information display assembled by Last.fm is proved more stable than initially expected. However, this stability does not last as a digital information display is, after all, a digital artefact which is always in constant flux. These dynamics lead to interesting implications for user choice, behaviour and demand for music on Last.fm.

Paper 3: How Does Data-based Music Discovery Work?

Here, we tested whether personalized music recommendations improve music discovery by building a statistical model for music discovery and music consumption of Last.fm users. Music discovery was considered a determinant of music consumption (path analysis). Results determined a significantly positive association between the use of recommender systems and music discovery; however, the magnitude of the positive association was relatively modest. This was especially so when we compared the magnitude with the negative impact of policy changes on Last.fm since 2009. From the inception of Last.fm to early 2009, users could stream free music directly. These free streaming services have now been terminated and users are encouraged to stream music provided by partners, such as Spotify. Music consumption declined directly as a result of this policy change. Consumption also declined indirectly because Last.fm users found it harder to discover music since the amount of behavioral data Last.fm collected was significantly lower. Besides these findings, a significant positive association was also detected between data quality and music discovery and between the level of social media user engagement and both music discovery and consumption.

This statistical model is also well-grounded in the existing literature. Bateman et al. (2011) show that online participation is directly linked to commitment as defined by organizational commitment theory. They identify three types of commitment as continuance, affective and normative. Music discovery enters as an explanatory variable for music consumption regarding continuance commitment, while social media user engagement enters as an explanatory variable for music consumption regarding effective/normative commitment. Uses of a recommender system and data quality enter as explanatory variables for music discovery. The former is obvious (Erkstrand et al., 2010), while the latter is related to how well Last.fm manages to counter problems stemming from inconsistent music metadata circulating in the digital ecosystem (Morris, 2012; Brookes 2014a, 2014b). Lastly, social media user

engagement also enters as an explanatory variable for music discovery as users can discover music by browsing through the listening profiles of their friends (Chen et al., 2010; Goldenberg et al., 2012).

Paper 4: More than Networks: Social Media as Infrastructure

Despite the growing importance of social media technologies for the current development of the web economy (i.e. social buttons, APIs, etc.), the majority of IS contributions continue to see social media platforms predominantly as social networking sites. The static models of network analysis cannot capture the dynamics of the layered architecture of data exchange that underlies the complex infrastructuring of social media. They consequently risk missing what constitutes the novelty and specificity of these platforms: the distinct ways by which they produce, circulate and commercialize data and the new forms of interaction they propose.

We conduct an empirical study of Last.fm, a social media platform for music discovery, and we find that social media technologies are strongly associated with user listening activity, which results instead only tenuously linked to community participation. Our study lends support to the view of social media as infrastructures resting on integrated and layered social technologies that filter social participation, sustaining a continuous flow of social data across infrastructure layers and (increasingly) across business domains.

The primacy of social media technologies as generative mechanisms of social media networks suggests that firms cannot view social media simply as a tool fostering community participation or engagement. We provide evidence of the importance of integrating an infrastructural approach to the partial view of social media as networks. We conclude by discussing the evolution of social media infrastructuring technologies and the making of the "social web".

Overall contribution

Systematic research on the operational aspects of social media

So far, research on the operational aspects of social media has been diversely scattered. Databases combine with algorithms in the construction of social media interfaces and the corresponding relations to user behaviors and how these are captured as social data and fed back into the system. This is very complex because the distinction between the consumption and operational aspects of social media is purely analytical. Consumption is captured as social data which is then fed back into the inner operation of social media demonstrating that the mechanism is really complex. Most papers study the production side of social media and focus only on one area of the inner operation as either algorithms, data or interface and user behaviors. The only paper that comes close to uncovering the whole picture of the inner operation of social media is written by van Dijck and Poell (2013) entitled "Understanding Social Media Logic". They juxtapose social media logic with mass media logic and highlight the distinguishing logical aspects underlying social media as programmability, popularity, connectivity and datatification. This thesis applies the theory of digital objects to elucidate the inner workings of social media which are conceptualized as composed of numerous modules (interfaces, databases and algorithms). This conceptualization allows the discussion of factors which can also affect components of the inner operation of social media, connecting the digital ecosystem with the database of social media platforms and business optimization with algorithms (see Figure 6).

Among the components which constitute the inner operations of social media, algorithms are the least studied. EdgeRank of Facebook has already been subjected to critical scrutiny (Bucher, 2012). However, a recommendation algorithm which is widely deployed on social media platforms has yet to be subjected to critical scrutiny. This thesis particularly concerns recommender systems which are discussed at length in the second section. While other papers

relating to the inner operations of social media refrain from studying algorithms and tend to treat them as opaque black-boxes, this thesis tackles algorithms head-on. To the best of my knowledge, this is the first piece of work that imports literature on recommender systems alongside literature on IF/IR which stems from technical/engineering literature. While some papers exist on information systems and media literature which discuss recommender systems their goal is not a critical evaluation but to propose new types of recommendation algorithms (Colace et al., 2015).

Recommendation algorithms can be critically assessed. There are basically two types of recommender systems. While IF can be considered as a subset of IR ("zero query" search), recommender systems can be considered as a subset of IF. Indeed, the two types of recommender systems follow the two paradigms of IF as the distinction made by Malone et al. (1987). One is a CB recommender system whereby item similarity is constructed according to product feature and the other a CF recommender system whereby item similarity is constructed according to ratings made by the user community. Each type of recommender system suffers from different problems. CF recommender systems have popularity bias while CB recommender systems constitute shallow text analysis. While both suffer from cold-start problems, this is more serious for CF recommender systems as they rely on the whole community of users to construct item similarity. Some authors believe that creating a hybrid recommender system would resolve these problems; however, one issue remains as the inability of recommender systems to recommend novel items. This occurs because the main tenet of both recommender systems is to "recommend similar items to those that users already like". Fortunately, this problem can be resolved by unleashing the full power of the third order as discussed by Weinberger (2008) which allows items to be sorted and retrieved in seemingly limitless numbers of ways, allowing for novel discovery.

Assessing the seven conjectures

Toward the end of the second section, seven conjectures derived from the theoretical framework are proposed as (1) navigation of information display as assembled by social media is highly interactive, (2) information organization of social media is highly unstable which would also render user behaviors unstable, (3) quality of data aggregation has significant implications on user behaviors, (4) the amount of data captured by social media platforms limits the usefulness of their information displays, (5) output from the recommendation algorithm (recommendation list) has real implications on user behaviors, (6) circle of friends on a social network can influence on the behaviors of users, and (7) metadata attached to items being displayed has influence for the behaviors of users. They have been assessed with either descriptive statistics or hypothesis testing by papers included in this thesis. In this section they are assessed individually.

Navigation of information display as assembled by social media is highly interactive

Information displays on social media are predominantly composed of ordered lists. This is because the display is constructed from behavioral data which is 'discrete and granular' and countable. Thus, a numerical value can be easily assigned to each item and then used to organize information assembled on the information display. Behavioral data can also be used to present users with a set of unordered items, but the belief is that ordered lists are more effective in assisting user discovery. This is why exact match and Boolean models are viewed as inferior to vector space and probabilistic models among classical IR techniques.

Seemingly limitless metadata can be attached to items in the digital information environment and this allows an infinite number of ordered lists to be assembled. The strategy followed to organize information in the digital environment is to include and postpone classification until an information need arises. This is dissimilar to recording metadata on catalog cards where

space is a restriction. The digital strategy is to sort on the way out, not on the way in. Take user-generated tags as an example. Seemingly limitless numbers of labels can be created and attached to items to be organized and it is simple to produce ordered lists corresponding to each tag. Labels attached to items need no longer follow, say, the Dewey decimal classification established by experts but can relate to anything including personal uses, desires, ambitions, impressions and moods.

Popularity ranking (ranking by consumption) can be assembled for each tag and prototype classification emerges out of this. An object no longer has to be located in a single place, it can be attached to seemingly limitless numbers of ordered lists. It might be located as one of the top items in one list but further down the list in another. Users can interactively browse these ordered lists according to their interests to discover items that they want. It is not necessary for ranking to be dependent on popularity as there are also other metrics. For example, items can be ranked according to the degree at which they are trending or by the growth rate of their consumption. One downside of ordered lists assembled by popularity is that they are likely to show items users are already aware of (such as Harry Potter books or music from Ed Sheeran). Ordered lists assembled by the growth rate of consumption may enable more novel discoveries as they are more likely to reveal less popular items. Adding more kinds of ordered lists into the digital information environment makes the information navigation experience even more interactive.

Last.fm, the empirical object of this thesis, utilizes all sorts of ordered lists to organize music. These include all kinds of music charts e.g. most popular artists/tracks by categories as assembled by user-created tags, trending artists/tracks by categories and most a ‘loved’ artists/tracks by categories. There exists a ‘love’ button on Last.fm whereby users can click as they listen to tracks of music, equivalent to the famous ‘like’ button on Facebook. These music charts are created weekly and users can also browse through charts created in previous

weeks. Further, users form their own groups and set their own objectives for those groups. Charts can also then be produced according to behavioral data retrieved from users within those groups. Charts entail only basic computations. Some ordered lists are constructed with more complex computation as lists of similar artists and lists of recommended items. While information navigation within a digital environment like Last.fm is highly interactive, it is also important to note that not all ordered lists are created equal. Some lists are featured more prominently on the interface and these are more likely to be browsed by users.

Information organization of social media is highly unstable which would also render user behaviors unstable

Ordered lists cast real implications on user behaviors because items higher up on the list are more likely to catch their attention and be discovered and consumed. This is the ranking effect of search cost to scroll down the list and browse through items which are ranked lower. Given that the databases of social media platforms are being constantly updated, interfaces assembled from the data are likely to be unstable. In other words, ordered lists assembled by databases are likely to be unstable. This is the nature of digital objects which are constantly in flux and lack stability. An unstable interface implies that the behaviors of users are also likely to be unstable because the interface has real implications through ranking effects. This is empirically demonstrated by Ghose et al. (2012).

It is true that interfaces cast real implications for user behaviors but are the interfaces really unstable? This very much depends on the algorithms used to assemble ordered lists from data or the kind of ordered list. Espeland and Sauder (2007) study USNews law school rankings and uncovered numerous self-reinforcing mechanisms. For example, student decisions are correlated with ranking, which also determines the quality of applications schools receive. Budgets can also become tightly linked to rankings as departments compete for funds with other departments within the same university. Furthermore, being lowly ranked makes it

more difficult to solicit alumni, making it even more difficult to generate revenues and resources to bump up the ranking. To analyze whether the information display of social media is unstable, the main question to ask is whether these self-reinforcing mechanisms exist for a particular display. If they exist, then it is likely that the display will be stable, if not then it is likely to be unstable.

Some ordered lists assembled by social media are likely to be stable and some are likely to be unstable. An example of the former is the popularity chart whereby a list is ordered according to the amount of consumption. Items higher up in the popularity charts are likely to be the usual candidates. Being higher up the popularity chart makes them even more popular and so this kind of information display is self-reinforcing and is likely to be relatively stable. This is different from a trending chart which is likely to be relatively unstable. It is easier for lesser known items to be a trending because it is easier to achieve high consumption growth, the condition needed to be on the trending chart. After all, trending charts were invented to ease the problem associated with the stability of popularity charts since charts are likely to be less useful to users if they keep on showing the same items.

Interestingly, this thesis determined that a similarity network assembled by a recommendation algorithm is self-reinforcing together with the operation of a recommender system. The main tenet of a recommender system is to "recommend similar items to those that users already like". Thus, similarity networks strengthen as users accept recommendations made by the recommender system. This thesis found that similarity rankings cast real implications on user behaviors and that they are relatively stable during business-as-usual operations of social media. Therefore, similarity rankings have the potential to stabilize consumption. However, the stability of similarity rankings and all ordered lists assembled by social media do not last forever. Constant business optimization is performed by social media whereby algorithms which assemble information displays are adjusted

leading to instability. This will eventually render the consumption of items being displayed and overall user behavior unstable.

Quality of data aggregation casts significant implications on user behaviors

Data aggregation has always been problematic in the digital ecosystem which produces data so large that it cannot be handled and standardized by experts within a single organization. Hence, data creation has become distributed resulting in the distributed creation of identifiers for diverse pieces of data. Two people may refer to an item differently and the same names can be imposed on disparate objects. Furthermore, even if people intend to refer to an object in the same way (by using the 'basic level' category), they may make spelling errors or use different spelling variations (e.g. adding brackets or tildes). While people can easily resolve these errors and variations it is not a simple task for a computer. Alas, it is important to rectify these inconsistencies because they impact on effective data organization in the digital information environment. It would be unhelpful for users to scroll down an ordered list and find the same item appear again and again but with slightly different name variations.

The problem of data aggregation is especially troubling for the operation of recommender systems. As already discussed above, one of the problems faced by recommender systems is cold-start problem whereby items receive insufficient rating which hinders similarity network construction and recommender systems are unable to recommend them to users. With the problems of data aggregation, rating data for an item is split into different lumps and each lump by itself may be insufficient for a reliable similarity network to be constructed. Alas, the problem is exacerbated when social media sites are transformed into social media platforms and made interoperable with the whole digital ecosystem. If rating data is only collected as users navigate and consume content within the same websites, name consistencies can be better ensured as labels within a single website that can be managed by a

single group of experts (webmasters). However, when social media also retrieve rating data from the digital ecosystem this is no longer the case as naming becomes a distributed activity.

Take the music industry as an example; currently, there are at least five metadata silos in the digital music ecosystem as *Gracenote*, *All Music Guide*, *Echo Nest*, *Discogs* and *MusicBrainz*. Each has different strengths and weaknesses, and standardizes music metadata in dissimilar ways. However, the problem of data aggregation is not restricted to the music industry. The 'like economy' also faces the same problem whereby similar items may be assigned with different names or identifiers. This is also true for the Quora and Stackoverflow websites where users can present any questions to the community. The same questions worded in different ways are being asked over and over again.

Resolving the problem of name inconsistencies is a never ending task as the activity of naming becomes distributed. Inconsistencies will continue to arise as users create social data. Nonetheless, it is important for social media to resolve these inconsistencies as much as possible and improve the quality of data aggregation. In the case of Last.fm, the association between quality of data aggregation is more strongly related to user behaviors than how the information is being organized (i.e. numerous ordered lists assembled on the interface of social media).

The amount of data captured by social media platforms limits the usefulness of their information displays

Social media requires sufficient data to assemble information displays for users. Popularity charts can be rendered unreliable if they are produced from data retrieved from a few users. Furthermore, as discussed above, cold start problems can be an issue for recommendation algorithms which can only be resolved by sufficient data for efficient operation of recommender systems. Also, data needs to be constantly updated. For example, if the latest

charts cannot be produced, users have to rely on older outdated charts for content discovery. Connection of social media to the wider digital ecosystem causes a data aggregation problem; however, this connection is important as it enables social media to maintain updated data organization. If social media relied only on behavioral data, derived from its user base to organize information displays, it would be impossible to capture new items which are constantly created, because the captured behavioral data would only be metadata concerning existing items already in the database. Retrieving behavioral data from the wider digital ecosystem injects newness into social media information displays and enables them to accommodate new items being constantly created.

The importance of social data for social media operation also implies that the social media user base generates positive externalities through the creation of social data. While the importance of the user base in terms of maintaining user interaction is widely recognized, the importance in terms of the creation of social data is given less priority. A large user base implies that there will be more social data created, leading to better information organization assembled by social media, which, in turn, generates an even larger user base. There are positive externalities created, by large user bases through creation of social data. If this cycle of growth can be triggered, then it is possible for social media to keep on growing and assemble better and better information displays. Unfortunately, the opposite can also be triggered whereby reduction in the size of the user base results in decline with less social data created and reduction in the quality of information displays.

This is what happened to Last.fm which achieved rapid growth of new user subscriptions until 2009 when the website decided to gradually wind down music streaming services and focus only on music discovery services. Last.fm encouraged its users to stream music from elsewhere such as Spotify. This upset many users and the growth of new user subscriptions declined, eventually leading to a reduction in the number of active users. Negative

externalities whereby a smaller user base leads to an even smaller user base through social data must be at work. Statistical models were fitted to user data before 2009 and from 2009 onwards and determined that the ability of users to discover music and the amount of music consumed per user declined sharply. This can be attributed to the negative externalities which triggered by the change of the business model of Last.fm.

Output from recommendation algorithm (recommendation list) has real implications for user behaviors

Recommender systems have been extensively employed by social media. So, how do they successfully shape user behavior? After all, recommendation algorithms only produce one of the ordered lists which can be browsed through by users. Some anecdotal evidence exists that recommender systems do shape the behavior of users e.g. 2/3 of movies rented by Netflix are recommended, recommendations generate 38% more click-through for Google News and Amazon claim that 35% of its product sales result from its recommendation engine (Celma and Lamere, 2007). Do recommender algorithms lead to increase in actual click-through and product sales or do recommender algorithms cannibalize sales through other channels (i.e. through other ordered lists)? To the best of my knowledge, this thesis provides the first answer to this question as recommendation algorithms do successfully influence behaviors of users and increase their consumption. In the case of Last.fm, the more similar the contents (according to similar network) consumed by users, the more overall contents were consumed by users.

This implies that the discovery process of users succumbs to the limitations imposed by recommender systems. First, a similarity list tends to rank popular items higher and so users are more likely to encounter this merchandise. Second, the main tenet of a recommender system is to "recommend similar items to those that user already like". Users will receive only safe and hardly novel recommendations. With these limitations, why are users still

willing to accept recommendations assembled by recommender systems? The answer is that safe recommendations are quite sufficient for some users; not all users want to make novel discoveries. Emap, a UK based company which owns several magazines and radio stations, carried out a study under the name Project Phoenix in 2003 looking specifically at the attitudes of people between the ages of 15 and 39 toward music (Jennings, 2006). Results show that 9% of users are identified as 'savant' or those where everything in life seemed to be tied up with music, 16% are identified as 'enthusiasts' with music as a key part of their life but balanced with other interests, 26% are identified as 'causal' with music playing a welcome role but other things far more important and 48% are identified as 'indifferent' who would not lose much sleep if music ceased to exist. Despite all their limitations, recommender systems can assemble good enough lists of music for the 'causals' and 'indifferents' who account for a large proportion of the population.

Circle of friends on a social network can influence the behaviors of users

Social networking functionality is often embedded into social media and a circle of friends on social networks help users to discover new contents. As friends of users discover and consume contents these are shown up in the newsfeeds of the users, so friends of the users also become exposed to the contents. Further, users can browse through the profile pages of their friends and see what items they have consumed. As already discussed above, the discovery of new items through social networks is very important as it helps to counter the limitations of recommendation algorithms. According to Goldenberg et al. (2012), profile pages of users have unique structural properties making them better content brokers than similarity networks as users are likely to post links to more variety of contents, bridging different product circles. As in the case of Last.fm, a small-scale survey by Chen et al. (2010) demonstrates that browsing through profile pages of friends is an important way for users to discover contents. This thesis further confirms that the number of friends is an important

determinant of the ability to discover new contents and the number of contents consumed by each user.

Metadata attached to items displayed casts influence on the behaviors of users

Limitless kinds of metadata can be attached to items to be organized and displayed. Content discovery is an important functionality of social media and profile pages are constructed not only for users but also for contents. Profile pages often exist for each item organized on social media and these profile pages may feature numerous metadata created by users and attached to the items. Do these metadata cast implications on the behaviors of users? This depends on the kind of metadata which is attached.

Content enriched metadata in bibliography records are helpful to library users as they try to discover relevant materials (Tosaka and Weng, 2011). Here, content-enriched metadata go beyond basic metadata (such as titles) to include content notes, summaries, tables of contents, simple text and other publication related material. However, this thesis determines that these content-enriched metadata do not have implications for user behaviors. This is because a quantitative analysis considers only the quantity of content and does not take quality into consideration (for instance the relevance of the content). However, this thesis considers that the role of user generated content on social media may be less important than previously assumed.

Another kind of metadata that can be attached to items to be organized are messages from users which culminate in chains of comments. It is surprising that chains of comments has implications for the behavior of users even though content enriched metadata do not. Perhaps, this demonstrates the social nature of content discovery and further supports the finding of the importance of the circle of friends in content discovery. Users may want to consume the same item because this allows them to socialize with one another. The majority of the

population can be considered as ‘causal’ and ‘indifferent’ types of content consumers who do not strive for eccentricity. These facets question the viability of long tail business models. Perhaps the long tail model is more imaginary than real as users are more likely to consume the same popular contents rather than items in the long tail (Elberse, 2008).

The last piece of metadata which can be attached to items to be organized is the content location. This piece of metadata is particularly important for social media which is aimed specifically at content discovery rather than content distribution. Last.fm is that particular kind of social media. There are also other websites which are built solely for discovery, not distribution. Examples include skyscanner.net and traveloka.com which compare air travel fares and hotel accommodation costs in different countries across diverse websites. Here, links to websites which distribute items being organized are particularly important as discovery will be deemed useless if users cannot get to their location. Indeed, connecting discovery platforms to distribution platforms is not an easy task given the constantly shifting digital ecosystem and URLs where the actual contents are located are always in flux.

Conclusions and future research

This thesis accomplishes three tasks. First, it carves out a relatively new area for information systems and media scholars to focus their research efforts on the operational aspects or the inner workings of social media. Research in this area is small but growing. While more than one thousand papers are identified which address the consumption side of social media research (i.e. why it is used, how it is used and what are the consequences), only one hundred are identified as covering the operational aspects. It is important and fruitful to study how social media operates because this shapes the behaviors of users in specific ways.

Second, this thesis built a theoretical model of social media information display. The theory of digital objects was applied and the operational aspects of social media demonstrated as the

three components of database, algorithm and interface. Data are collected as users participate in social media and also accrue from the broader digital ecosystem as social media sites transform into social media platforms. Algorithms operate on the databases of social media sites to assemble interfaces for user interaction. To the best of my knowledge, this is the first paper in the field of information systems and media studies which critically assesses recommendation algorithms. While such algorithms have been widely deployed they have yet to be critically examined, unlike other well-known algorithms such as the EdgeRank of Facebook. Results show that all recommendation algorithms suffer from one shortcoming as the inability to recommend novel items. Fortunately, this problem can be resolved by unleashing the power of information organization in the third order and especially by browsing through profile pages of other users.

Third, this thesis presents seven conjectures derived from the theoretical model which are assessed against analysis of data collected from Last.fm. This analysis supported most of the conjectures except the stability of information display and the importance of metadata attached to items being displayed. The former is interesting because the theory of digital objects posits that they ought to be in constant flux. However, this is not necessarily the case for numerous ordered lists assembled by social media and used to display information and depends on whether a self-reinforcing mechanism exists. If it does, then that particular ordered list will exhibit stability rather than instability. However, it is important to note that such stability is not permanent because algorithms can always be fine-tuned as social media optimize their business operations. As for the latter, not all metadata attached to items being displayed influence the behaviors of users. Interesting, user-generated content enriched metadata do not cast implications on user behaviors. This demonstrates that the role of user content generation on social media may be less important than previously assumed.

Further studies can address some limitations of this thesis along several directions. First, it is extremely relevant to point out that the quantitative research design employed here does not, in any way, account for the causality. A quantitative model by its very nature accounts only for correlation by measuring association among variables. The statistical models in this thesis provide rich evidence to prove the strong association between various variables of interest. Having said that, the statistical models assembled here have their own limitations and more diverse empirical evidence is needed to investigate the construction of information displays. This can be done by tapping into primary qualitative data collected through interviews or participant observations. Second, this is the first piece of work which uses the theory of digital objects to conceptualize the construction of information displays by social media. Last.fm is used as a prime empirical object. Much work is still required to study information displays of other social media platforms to determine whether the conjectures stipulated in this thesis hold up against new empirical evidence.

Literature Review of the Production Side of Social Media

Akarapat Charoenpanich, LSE

Abstract

This paper presents a systematic review of the literature on the production side or the operational aspects of social media or its inner workings behind the user interface. It examines eight information system journals (Senior Scholars' Basket of Journals) and ten media journals related to technology between 2000 and 2016. A total of 1,732 papers focused on the use of social media or the consumption side with only 99 papers identified on the operational aspects which comprise a relatively small but growing area. The number of papers published on the consumption side began to gain traction by 2007 and accelerated further in 2013. However, research on the operational aspects has grown steadily since 2013 and presents a fruitful area for further studies. The inner workings are of importance for the operation of social media as they cast real implications on users' behaviors.

Papers on the operational aspects were then furthered classified into four categories as algorithms (16 papers), data (26 papers), interfaces (54 papers) and reaction against the commercial aspects of social media (3 papers). Predictably, there were fewer studies on algorithms than interfaces; the latter are more visible while the former are seemingly invisible and are often viewed as opaque or 'black boxes' by information systems and media scholars. Although algorithms have been widely studied by computer scientists, they often conduct research in a technical manner and, hence, do not uncover the societal implications of information systems which are deemed important by media scholars. Papers concerning the operational aspects in each category are then reviewed with future research directions discussed.

Introduction

Social media has become increasingly important in everyday life. In June 2017, Facebook claimed two billion active users representing 27% of the global population (7.5 billion). A systematic review of the operational aspects or production side of social media is undertaken to better understand the implications of user behavior. This paper is divided into five parts. The first asserts the lack of detailed research on the subject while the second discusses the methodology employed along with the definition of social media. Papers concerned with peer productions (e.g. open source projects), multiplayer online games and virtual worlds as empirical objects are excluded. 18 journals covering information systems and media research from 2000 to 2016 are included. The third part discusses the quantitative findings, demonstrating that the research on the production side of social media is still relatively small as compared to the consumption of social media, and the fourth reviews the literature regarding the operation or production side of social media which are divided into four categories as data, algorithms, interfaces and reaction against commercial social media. The final part considers aspects for future research.

Review of existing literature reviews on social media

There are many literature reviews of social media; however, none of these explicitly assesses the operational aspects. One of the most influential reviews is written by boyd and Ellison (2007) entitled 'Social Network Sites: Definition, History, and Scholarship'. This paper plays an important part in establishing social media as a field of study and defines social networking sites as "web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection and (3) view and traverse their list of connections and those made by others within the system". Interestingly, boyd and Ellison (*ibid.*) distinguish social network

sites from social networking sites; the latter are viewed as a subset of the former. They suggest that networking (meeting with strangers) is not the primary practice of many social media users who merely communicate with people who are already a part of their social network offline. This demonstrates the user-oriented nature of their research agenda which classifies social network sites as merely there to support already existing offline relationships. In general, the detailed operation of social network sites does not enable individuals to forge new online connections or interfere with user behavior. They also discuss four research agenda: (1) impression management and friendship performance, (2) network and network structure, (3) bridging online and offline networks and (4) privacy. All of these facets can be considered as the consumption side of social media.

Kane et al. (2014) modify the meaning of social networking sites as proposed by boyd and Ellison (2007). They highlight four features shared by social media technologies as digital profile, search and privacy, relational ties and network transparency. The latter two features are similar to the definition of boyd and Ellison (*ibid.*) while the former two are not. Digital profiling extends beyond exclusive intentional and conscious construction to automatic and passive records of user activity. People can also access content on social networking sites without directly viewing digital profiles. For example, streaming content can be automatically filtered and users might engage in search activities for keywords in LinkedIn profiles to find people with particular skills or experience. The ability to search for contents has raised privacy concerns and data protection has now become a significant social media issue. Most social media sites offer features to control access to user contributed content. Kane et al. (2014) attempt to discuss the inner working of social media but leave many questions unanswered; they contend the idea of a bounded system embedded in the definition of boyd and Ellison (2007) but do not incorporate this into their new definition which still focuses on the consumption side of social media.

Zhang and Leung (2015) review the literature on social networking services in six major communication journals between 2006 and 2011. They discover that most papers can be allocated into the four themes discussed by boyd and Ellison (2007). In addition, they reports demonstrate how trust, attraction, emotional closeness, emotional support and perceived social support are facilitated by social networking services. Many studies adopt a psychological approach, incorporating intrapersonal traits such as self-esteem, collective self-esteem, happiness, satisfaction, emotional openness and extraversion. In general, past research reports that people experience increased happiness and excitement through the use of social networking sites. Nonetheless, some personality characteristics negatively affect offline and online communication including loneliness, jealousy, communication apprehension, narcissism and neuroticism. Zhang and Leung (2015) point out that future research should emphasize the role of networks, improve ease of use, rethink the nature of relationships and friendships on social networking sites, consider the dynamic adoption process and expand to cross-contextual and cross-cultural domains.

Rains and Brunner (2014) review research related to social networking services published in six interdisciplinary journals between 1997 and 2003. They determine that over 66% of the studies are limited to a single company and 80% of these explicitly examine Facebook. Microblog studies are even more concentrated with over 90% in the six journals limited solely to Twitter. These results concur with Zhang and Leung (2015) and Zhang et al. (2015a) who point out the importance of Facebook. Zhang et al. (*ibid.*) searched for papers on social media in the Citation Databases under 'business and economics' and 'computer science' and suggest that studies on social media in different disciplines are not well-combined. They attempt to obtain a more complete picture by investigating the discipline of management and computer science. Their data indicate that the most cited study is by Kaplan and Haenlein (2010) entitled 'Users of the World Unite! The Challenges and Opportunities

of Social Media' which explicitly and systematically defines social media. Further, the largest bibliography is attained by the classic paper of Granovetter (1973) entitled 'The Strength of Weak Ties'. While most cited papers on the business side concern word of mouth communication, papers in the field of computer science mainly cover analytical techniques (topic modeling and social network analysis). Studies of social media increase rapidly after 2009 with Facebook emerging as one of the top keywords (Zhang et al., 2015).

So, how do authors study Facebook? Caers et al. (2013) assess scientific, peer-reviewed articles on Facebook between 2006 and 2012 extracted from the Institute for Scientific Information (ISI) Web of Knowledge. Thematic topics covered in their corpus include initial motivations to join Facebook, characteristics of Facebook users, building and maintaining a Facebook network, motives for disclosing information on Facebook and the effect of disclosing this information. They review papers concerning the organizational uses of Facebook including how the social networking service attract customers and future staffs and identify numerous gaps in previous studies including why former users decide to abandon Facebook, cyber-bullying and the extent to which information disclosed by users reflects their actual personality traits, motivation and competence. Wilson et al. (2012) classify papers on Facebook into five categories as the descriptive analysis of users, motivations for using Facebook, identity presentation, the role of Facebook in social interactions and personal information disclosure.

However, none of these reviews focus on the inner working of social media operations and how that, in turn, shapes user behavior which they do not regard as a fruitful area of research. Numerous literature reviews also address other limited features of social media but none concentrate on the operational aspects. Sun et al. (2014) and Malinen (2015) examine the literature on social media user participation. Zhang et al. (2015b) and Leonardi et al. (2013) review the use of social media to support knowledge management, especially in business

enterprises. Gallagher and Savage (2013) assess the literature regarding the cross-cultural analysis of social media while Li et al. (2014) focus on social media usage in China. Oinas-Kukkonen et al. (2010) review literature on network analysis applications for social media while Kumpel et al. (2015) examine social media news sharing. Khosravi et al. (2016) investigate the social media impact on the loneliness of senior citizens while Baker and Algorta (2016) and Moreno et al. (2016) look at the relationship between social media, depression and alcohol consumption.

Methodology

This paper follows the definition of social media adopted by Kaplan and Haenlein (2010) as "a group of internet-based applications that build on the ideological and technological foundation of Web 2.0 and allow the creation and exchange of user generated content". They identify six categories of social media as blogs, collaborative projects (e.g. Wikipedia), social networking sites (e.g. Facebook), content communities (e.g. YouTube), virtual social worlds (e.g. Second Life) and virtual game worlds (e.g. World of Warcraft). To ensure that social media aspects reviewed in this paper relate to everyday public usage, collaborative projects (such as open source software production), virtual social worlds and virtual game worlds are excluded. Dating services such as Tinder and online communities including Slashdot and Reddit are included as they build on "the ideological and technological foundation of Web 2.0".

Papers in 8 journals covering information systems (Senior Scholars' Basket of Journals) and 10 journals in the media field are reviewed between 2000 and 2016 (if possible). The former include:

- European Journal of Information Systems
- Information Systems Journal

- Information Systems Research
- Journal of AIS
- Journal of Information Technology
- Journal of MIS
- Journal of Strategic Information Systems
- MIS Quarterly

The latter include:

- The Information Society
- Journal of Computer-Mediated Communication
- Computers in Human Behavior
- Cyberpsychology, Behavior, and Social Networking
- New Media & Society
- Surveillance and Society (established in 2002)
- Computational Culture (established in 2011)
- Media and Communication (established in 2013)
- Big Data and Society (established in 2014)
- Social Media + Society (established in 2015)

Articles were collected between 2000 and 2016 from all the journals listed above except for Surveillance and Society, Computational Culture, Media and Communication, Big Data and Society and Social Media + Society as these were established after 2000. Only research articles, commentaries and teaching cases were included in the corpus while editorials, introductions to special issues, obituaries and book reviews were excluded. Titles, abstracts and keywords were extracted from the articles and concatenated into a single text. Each text was then screened for the words, 'social media', 'social network', 'Web 2.0', 'online communities' and 'online community' plus the names of social media and online dating sites

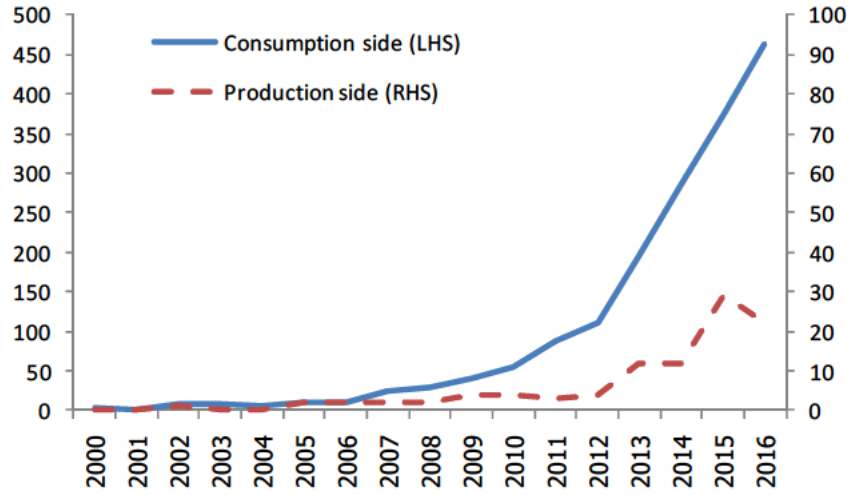
identified in Wikipedia. If the words and names were not found, the texts were excluded from the corpus. Each retained text was screened to determine its relationship to social media as defined in this article. The total corpus consisted of 1,854 papers. When papers on political economy were excluded, the size of the corpus was reduced to 1,831 papers constituting the literature relating to the consumption and production side of social media.

Quantitative analysis

Out of these 1,831 papers, 1,732 were classified as papers on the consumption side of social media, while 99 were classified as papers on the production side of social media. Discussion of the consumption side of social media has already been reviewed in the first section of this paper regarding why social media usage is mainly linked to psychological traits and motivations, and how social media is used with emphasis on the diverse nature and consequences of such uses, both utopian and dystopian. Figure 1 shows the number of papers regarding the consumption and production side of social media through time. Papers published on social media consumption started to gain traction in 2007 and increased rapidly after 2013, while research on the production side also grew steadily after 2013. The latter is a relatively small but important growth area with potential for fruitful research in the future. The next section reviews the 99 papers collected on the production side or operational aspects of social media.

Figure 1: Number of papers on the consumption and the production side of social media

Unit: Paper count



The 99 papers on the operational aspects of social media can be further classified into four categories as algorithms (16 papers), data (26 papers), interfaces (54 papers) and reaction against commercial social media (3 papers). These components underlie the operation aspects of social media and they will be explored further below. Users interact with an interface assembled by the social media software, which, in turn, shapes their behavior. This software component is composed of an algorithm and a database. Algorithms or computational procedures operate on data stored in databases to assemble interfaces for user interaction. Among the small but growing literature on the operational aspects of social media, algorithms are less frequently studied than interfaces because the latter are visible, while the former are seemingly invisible and are often viewed as opaque or 'black boxes' by information systems and media scholars. Although algorithms have been widely studied by computer scientists, they often conduct research in a technical manner and, hence, hardly uncover the societal implications of information systems which are deemed important by media scholars.

The production side of social media: algorithms, data, interfaces and reactions against commercial social media

Data are fundamental for social media operation. van Dijck (2013) states that social media logic is well-grounded in the condition of datatification which can be referred to as the ability of social media to "render into data many aspects of the world that have never been quantified before". Likewise, Kallinikos and Constantiou (2015b) cite Gillespie (2014) who argues that "algorithms are inert, meaningless machines until paired with databases upon which to function. A sociological inquiry into an algorithm must always grapple with the databases to which it is wedded". Therefore, this section starts with a literature review of social media data and subsequently addresses the literature on algorithms, interfaces and finally reaction against commercial social media. The last section is interesting as it sheds some light on the typical problems of social media setups and underpinning operations.

Data

Facebook is an exemplar of social media and its mission is "to make everything social" which actually means "to move social traffic onto a networked infrastructure where it becomes traceable calculable and so manipulable for profit" according to Couldry and van Dijck (2015). Andrejevic (2011) describes the role of online monitoring in the exploitation of user-generated activity. However, to be able to do this, Facebook requires a data infrastructure which encompasses various open source technologies such as Apache Hadoop, Hadoop Distributed File System (HDFS), MapReduce and Hive. Apache's HDFS is a popular open source distributed file system designed to meet the large demands of batch processing; MapReduce is a main component of Facebook's data analytic engine and Hive is a petabyte scale data warehouse built on the Apache Hadoop system (Van der Vlist, 2016).

Social media can host its own physical data infrastructure and choose to operate in the cloud with many data centers build to provide cloud computing services. An exemplar center is the Utah Data Center which gained media attention through the surveillance revelations of Edward Snowden. Data centers generally occupy large amounts of space, often equivalent to several football fields, are located in small rural towns and consume a lot of resources. The Utah Data Center covers a 1.2 million square feet enclosure situated next to 250 acres of sagebrush, so more storage can be added in the future. It can store data at the rate of 20 terabytes (information equivalent to the Library of Congress) per minute. As well as data collection and storage, the center also processes and correlates information on an ongoing basis. In 2013, the operational cost was US$ 1 million a month. Large companies often remain secretive about their infrastructure and estimations of how many data centers exist vary greatly. Emerson Network Power estimated that there were approximately 500,000 data centers in 2011 while The Register claimed 3 million as a more accurate guess in 2012 (Hogan, 2015).

But what are the characteristics of the information stored in these data centers by social media companies? According to van Dijck (2014), dataism is the widespread belief that social data are an objective qualification of human behavior and online sociality whereby social media are merely neutral facilitates. However, Shaw (2015) argues against this and points out that the process of generation shapes social data. For example, when the number of photographs uploaded per user of Flickr is plotted, spikes emerge at regular intervals because Flickr allows users to upload photographs in batches of six. Delicious has auto-completion in place which dictates how users assign tags to each item. An auto-completion system can be designed in various ways. For example, it might suggest tags most often assigned by other users to the same items or it might suggest tags often assigned by particular users who assigned tags in the past to others.

There are two kinds of social data. On the one hand, a kind of social data is structured data. Alaimo and Kallinikos (2016) point out that social media organizes online participation through stylization of social interaction and far-reaching standardization of online activities. This turns online behavior into discrete and granular data such as 'liking', 'tagging', 'sharing' and 'following'. There are many examples of these so-called behavioral data. Poor (2005) discusses social data used to govern Slashdot, an online community for computer enthusiasts, whereby each post has a score ranging from -1 to 5 according to how it is being monitored and each user has their own rating (Karma) according to how their posts are being rated.

Helmond (2013) discusses Uniform Resource Locator (URL) shortening services which have been popularized by Twitter with its limit of 140 characters on each tweet. Many social sites have their own specific URL shortening services such as flic.kr for Flickr, youtu.be for YouTube, fb.me for Facebook and t.co for Twitter whereby sharing contents may automatically produce shortened URLs. Numerous behavioral data can then be produced as users interact with these shortened URLs, for example number of clicks and date/time at which the links were clicked. Furthermore, aggregated data for all related shortened URLs can be assembled to produce a long URL. This provides social media with valuable information about popularity and spread of content which can be used to power trend topics and personalization features on their sites.

Biven and Haimson (2016) discuss gender which is another important aspect of structured data for social media advertising and marketing operations. Normally, the signup page of a social media site blocks access until all mandatory fields are filled. These mandatory fields often include binary gender; therefore, people with complex gender issues need to misclassify themselves to enter the site. Recently, social media has adapted to allow for non-binary genders. For example, Facebook added 56 custom gender options in 2014 while Google+ and Pinterest incorporated open text fields for users to enter any label they wished. However,

non-binary genders only occupy certain parts of social media sites as public profile pages and newsfeeds. Often, users are forced to select their preferred gender pronoun (he, she or they) which is converted back to binary gender to inform advertisers and marketers. Although Pinterest does not include a mandatory gender pronoun categorization, its strategic alliance with Facebook enables it to store binary gender data. Oakley (2016) studies gender construction on Tumblr and discovers that free-form labeling practices are still "grounded in hegemonic female/male, feminine/masculine binary discourse" and that "it is nearly impossible to fully break away from the dominant discourse".

On the other hand, another kind of social data is unstructured data which includes image, sound and visual records. They are difficult to categorize using traditional statistical techniques. For example, although digital images can be reduced to bits and dealt with computationally, their meaning cannot be easily controlled or manipulated (Constantiou and Kallinikos, 2015a). Unstructured data such as images can be more easily dealt with when they are combined with behavioral data such as tags which are another important and widely used behavioral data source. Tagging attaches a specific set of labels to objects and folksonomy (or folk taxonomy) emerged as a way to structure and share information (Derntl et al., 2011). Gehl (2011) points out that social media employs their users as information processors. For example, Digg users sift through massive amounts of digital information and rate them, allowing Digg to sort and organize digital information on the internet.

Importantly, Constantiou and Kallinikos (2015b) state that "data generation is lifted out of the prevailing expert-dominated cultures by which the information needs of practice fields have been defined and data are collected and stored" as "the outcome of the fundamental fact of making online interaction and the activities of large, shifting, heterogeneous and dispersed populations of users (mostly lay people) the drivers and carriers of data generation". Similarly, Pletrobruno (2013) analyze the transmission of intangible UNESCO heritage

videos on YouTube and determined that contributions of lay people countered the official heritage narrative, while Zervas and Sampson (2014) analyze the implications of tagging digital educational resources and suggested that tagging by lay people can enlarge relevant metadata compared to tagging performed exclusively by experts.

Constantiou and Kallinikos (2015a) identify four characteristics of social data. First, it is heterogeneous and often useful when amalgamated in an aggregation which uncovers context generality rather than a specific and contingent character of each data point. Second, it escapes the systematic nature of professional classification. Third, it crosses the border of alphanumerical systems and includes varying cultural artifacts cast in the media of text, image and sound, while finally, it requires constant renewal and updating. Users manipulate social data to construct their own identity and this is made possible because each user is assigned with permanent and persistent labeling. Deep profiling and the availability of past interaction archives allows users to learn about the identity of other site participants (Ma and Agarwal, 2007).

Data management is another important study area. Health technology has recently burgeoned with an estimate of 100,000 health applications listed in Google's Play Store and Apple's App Store (van Dijck and Poell, 2016). PatientsLikeMe is a real time health-related social media platform whereby users can upload and track their medical conditions (diseases, symptoms and treatments) and find patients with similar conditions patients who can offer support by sharing their own experiences. Self-reported data are then exploited for scientific and commercial medical research. PatientsLikeMe has produced 37 scientific publications based on data contributed by more than 220,000 patients (Tempini, 2015). This method of data collection differs from how data are collected for medical research as it is self-reported and does not rely on clinical interviews performed in institutional hospital environments by doctors and nurses (Kallinikos and Tempini, 2014).

Tempini (2015) discusses the data management challenges faced by PatientsLikeMe with conflicting demand for local context flexibility and the richness of data specificity. All patients differ from each other (they might have diverse levels of medical literacy) and each needs to be treated as an individual to enhance engagement on PatientsLikeMe. For example, patients can request the creation of new medical entities or definitions that are not available in the database. However, data created by users may not be deemed specific enough for medical research. In this case, local context flexibility may be limited to enhance data specificity; It might be necessary to differentiate between patients suffering from taxonomically close conditions (subtypes of the same parent condition). In this case, PatientsLikeMe may allow users to input only subtypes in the system, but not the parent condition. While this increases data specificity richness, some patients may not recognize the subtypes and overall engagement is dampened.

But, is this the best way to collect social data? There might be ways to achieve both context flexibility and data specificity richness. According to Parsons and Wiersma (2014), this is possible if one adopts instance-and-attribute based data collections rather than class-based conceptual models. Users usually accurately classify an instance at only 'basic level' in both free-form and schema-mediated data collections. This 'basic level' is widely accepted in cognitive psychology as the generally preferred classification level for non-experts. This is an intermediate taxonomy level (e.g. bird is a level higher than American Robin, but a level lower than animal) and is often the first class people think of when they encounter an instance. However, classification at 'basic level' implies a loss of information or data specificity richness and this is a problem of class based conceptual models. As an alternative, people may postpone classification and only report attributes of instances in instance-and-attribute based data collections. For example, 'standing on the ground' or 'orange back' can be attributes assigned to a bird. Once several attributes are reported for an instance, the

computer can then match them with pre-existing sets of identifying attributes and infer a class for that instance.

Lastly, some authors discuss data ecosystems created by social media. Helmond (2015) notes that social media sites are transformed into social media platforms when they establish application programming interfaces (APIs), rendering the platforms reprogrammable by third party developers. APIs were initially implemented as business-to-business solutions for e-commerce, enabling transactions and sales management. For example, Salesforce established APIs in 1999, eBay in 2001 and Amazon in 2002. In the mid-2000s, social media sites started to establish their own APIs. For example, Delicious established APIs in 2003, Flickr in 2004, and Last.fm, Facebook and Twitter in 2006. Developers can access platform data and functionality through APIs, enabling them to read, write and delete user data. Dissemination of the so-called widgets as plugin modular components enables integration of platform content and functionality into another website using a few lines of code. This includes social plugins such as the like button developed by Facebook. Technically, these social button functions as APIs call and send specific requests to Facebook's platform, for example, to ascertain the number of people who like the post or to publish the likes on the user's timeline. Alas, APIs can change as the business models of social media change. In the past, Twitter had a reputation as a data accessible platform as the Twitter API allowed easy scrape or download of massive amounts of data. However, Twitter imposed a download restriction of only 1% of traffic in 2011 and encouraged users to purchase data through a Twitter reseller such as Gnip (Felt, 2016).

According to Gerlitz and Helmond (2013), this has resulted in the 'like' economy. Facebook eventually extended beyond the limit of its platform and offered widgets which can turn websites and applications into a part of its platform. Social graph is an important component of Facebook as the representation of people and their connections to other people as well as

objects within the platform. In April 2010, Facebook launched Open Graph Protocol, which allows external websites and applications to be integrated with Facebook's Social Graph. Currently, more than 7 million applications and websites are integrated with the platform. Social plugin allows users to engage with content outside the platform through Facebook based activities such as liking, sharing and commenting. Once users click on the like or share button attached to external contents, these then become available for further liking and comment within the Facebook platform, generating additional data flow back to the external counter. Furthermore, data flows back to webmasters in the form of Facebook Insights with, for example, reports on the basic demographics of likers such as age, gender and location. Hence, webmasters are happy to grant Facebook real estate on their webpages in exchange for user engagement and Facebook Insights. Applications can also be integrated with Facebook's Open Graph. Using an eReading device called Kobo as an example, Facebook registers when users start reading a book on the device and will inform the user's friends on their newsfeeds when this happens. Chains of comment/like then follow these announcements which can be tracked by Kobo staff (Kaldrack and Rohle, 2014). In the same vein as webmasters, third party developers are happy to integrate their applications with Facebook in exchange for user engagements and insights.

Algorithm

"Social media platforms don't just guide, distort and facilitate social activity, they also delete some of it. They don't just link users together; they also suspend them. They don't just circulate our images and posts; they also algorithmically promote some over others. Platforms pick and choose"

Gillespie (2015)

So, "platforms pick and choose" according to Gillespie (2015) and the underlying algorithm appears invisible. Beer (2009) considers the aspects of software 'sinking' into and 'sorting' our everyday lives and cites Thrift (2005) who suggests that "software has come to intervene

in nearly all aspects of everyday life and has begun to sink into its taken-for-granted background". Other scholars also express similar concerns. Baym (2015) points out that opaque algorithms filter what one sees; users can neither understand nor influence these filtering mechanisms or comprehend the interest they serve. Sandvig (2015) investigates the secret process that determines relevance, judging whether something will be shown at all. For example, Facebook evaluates user generated content and may decide not to accept some of the posts. Braun (2015) notes that mechanical editors exist in Facebook, deciding algorithmically which posts and topics warrant inclusion from the continuous and often overwhelming feed of information delivered to users. Bucher (2015) mentions that users do not simply write articles and make their networks visible to others as networks are also articulated by underlying software and algorithms. Lastly, Shah (2015) suggests that information is communicated in social media mainly between machine and machine and not by humans. Similarly, Beer (2009) cites Hayles (2006) who claims that in "highly developed and networked societies … human awareness comprises the tip of a huge pyramid of data flows, most of which occur between machines".

van Dijck and Poell (2013) deconstruct social media logic. They assert that algorithms can steer users' contributions and shape all kinds of activities such as liking, favoriting, recommending and sharing. This has culminated into automated connectivity of users to content, users to users, platforms to users, users to advertisers and platforms to platforms. For example, Facebook and Linkedin present users with lists of 'people you may know', Flickr presents users with 'groups you may be interested in' and Amazon recommends items as 'people who bought this item also bought'. Compared to mass media, algorithmic assessment of information has replaced reliance on accredited experts and scientific evidence. Algorithms now have the ability to boost the popularity of people, things or ideas. Facebook's EdgeRank and Twitter Trending topics have the ability to promote certain

contents over others. Each social media also creates its own popularity metrics and tries to make them meaningful in social life offline. This includes view statistics for YouTube, friend statistics for Facebook and follower counts for Twitter.

Social media abounds with metrics. Grosser (2014) defines metrics as "enumerations of data categories or groups that are easily obtained via typical database operations and represent a measurement of that data". Metrics rely upon perhaps the most basic algorithm (summation) and are arguably the building blocks of more complex algorithms such as Facebook's EdgeRank, targeted advertisements and numerous recommendation and matching systems. Facebook is filled with metrics such as number of likes, comments, shares, friends, mutual friends, pending notifications, events, friend requests, message waiting, chats waiting, photos, places and much more. Facebook also produces metrics which are unseen by users, such as how many objects users like per hour, how many advertisements they click and the effectiveness of 'people you may know' at getting users to add more friends. So, how does Facebook choose which metrics to reveal to its users? The primary criterion is whether a particular metric will increase or decrease user participation. For example, users are more likely to click on an advertisement if they see that many people already like the object. Indeed, Facebook has the status of perpetual beta, whereby hundreds of experiments on small design variations and features are constantly rolled out. The impacts of alternative designs are compared and the most efficient at fostering user participation are selected. Users are unaware that they are the subjects of these tests and they have no choice but to steer toward the most efficient designs (Heyman and Pierson, 2015).

According to Bucher (2012), Facebook's EdgeRank is a powerful source of visibility on the Web; however, it has not been critically scrutinized. Therefore, Bucher (*ibid.*) investigates this aspect. Newsfeeds make up the central experience of Facebook users, representing constantly updated lists of posts from friends and interrelated pages divided into two areas.

One is Top News which aggregates the most interesting contents (according to EdgeRank) from friends and the other is Most Recent which shows all the actions of friends in real time. Every item shown in the newsfeed is considered an object and interactions with objects (like, comment, etc.) create what Facebook calls edges. The EdgeRank algorithm determines the content shown on users' Top News by drawing on different factors relating to edges. There are three components of EdgeRank. First is the affinity or the relationship between the viewing user and the item's creator. Sending a friend a private message or checking his/her profile on a frequent basis heightens users' affinity scores to that particular friend. EdgeRank assumes that users are not equally connected to their friends and some friends count more than others. Second is weight, whereby each edge is given a specific weighting depending on how important Facebook considers it to be. Not every edge is weighted the same as some types of interaction are considered more important than others. For example, comments are considered more important than likes. Third is time decay. This gives value to the freshness of the edge whereby older edges are considered less important than new ones. Higher ranked items according to the EdgeRank algorithm are more likely to appear in users' feeds and the weight given to certain edges depends on the internal incentives of Facebook at a particular time point. If Facebook wants to promote a certain product, say the 'Question' feature, then interaction with these features will probably be ranked higher than others.

Bucher (*ibid.*) highlights the discrepancy between what users think they should be seeing and what Facebook presents. In February 2011, Facebook changed the default setting for Most Recent feed to 'Friends and pages you interact with the most'; however, most users still believe that it represents every update from all of their friends in real time. Users were not notified of this change and the option to change the default setting is tucked away at the bottom of a drop-down menu. Bucher (*ibid.*) observes that the EdgeRank algorithm creates a threat of invisibility on the part of the participatory subject. To become visible, one would

need to follow the logic embedded within the EdgeRank algorithm and as such a whole new industry emerged around so-called 'newsfeed optimization'. EdgeRank does not automatically impose visibility on all subjects, thus, visibility is not something ubiquitous, but scarce.

To confirm this, Bucher (*ibid.*) conducted an experiment over a two month period from March to April 2011, comparing content in Top News to that of Most Recent feed and determined that only 16% of possible stories made it to the Top News. Further, a story published within the last three hours has a 40-50% chance of getting into the Top News feed. Subsequently, the threat of becoming invisible influences the actions of Facebook users who participate more to avoid disappearing and becoming obsolete. Highlighting posts with a lot of likes and comments creates the impression that participation is a norm and users who are bombarded by posts selected by EdgeRank are also likely to participate. Lastly, users strive to be popular on Facebook because popularity enhances the probability of becoming visible and thus generating even more interaction.

Birkbak and Carisen (2016) attempt to justify the EdgeRank algorithm which has been widely criticized as acting like a mechanical editor choosing what users see. However, perhaps this is inevitable given the huge amount of information being channeled at users. Without the EdgeRank algorithm in place to filter information, users will face the problem of information overload and might not be able to make sense of all the data. The EdgeRank algorithm has also been criticized as acting as 'echo chambers', 'filter bubbles' or 'walled gardens' whereby users receive only messages in conformity with their points of view. To counter this criticism, data scientists at Facebook argued that many weak ties exist in Facebook which help to spread novel information, demonstrating that social media can act as a powerful medium for sharing new ideas. Finally, the EdgeRank algorithm has been criticized because it can be gamed. In fact, a whole new industry focused on 'newsfeed

optimization' has emerged with the increasing utilization of Facebook. It can be argued that the EdgeRank algorithm cannot be easily gamed. To be granted with visibility, businesses need to produce relevant content for people who matter most. Furthermore, the depth of engagement matters as comments have more value than likes. Thus, businesses need to stop chasing algorithms and foster engagement in meaningful ways. This business advice reads more like a recipe for productivity rather than algorithmic tricks.

Targeted advertisement is another product of social media algorithms. On the one hand, Heyman and Pierson (2015) identify money as another factor that enters into the EdgeRank algorithm. Paid solutions exist to the threat of invisibility as discussed by Bucher (2012). Nonetheless, Facebook needs to strike a balance between affinity and profitability as users can be scared off by irrelevant advertisements. One of the products that Facebook provides for businesses is Sponsored Story (SPS), whereby posts of friends which are related to a specific page, application or other item advertisers want to promote achieve higher ranking in the EdgeRank algorithm and so are more likely to appear in the newsfeeds of users who are being targeted according to their interests or profile details. Here, the advertisement is camouflaged as user generated content. Although Facebook no longer provides SPS which was replaced by separate advertising services in 2014, the basic idea behind SPS remains.

On the other hand, Villard and Moreno (2012) investigate the numerous problems of advertisement systems on Facebook. They create fitness-related posts on Facebook accounts of college students and find that the fitness related advertisements that appear thereafter are irrelevant; some promoted products deemed too expensive for college students such as workout gear, while others are not accessible on the college campus such as charity runs across various states. They also determine that Facebook generates fitness related advertisements which may be harmful to the health of college students. For example, advertising for Fat Burning Finance aimed to lose weight was sent to underweight college

students. This happens because the Facebook advertisement system deprives targeted keywords of their context. Examples of messages uploaded by underweight college students include "was turned away from giving blood today because my weight was not enough". Perhaps, the advertisement was sent inappropriately because it targeted the word 'weight'. Similarly, advertisements for sweet and other junk food might be sent to overweight college students who post messages like "starting my diet today, no more chocolate for me!", because of the word 'chocolate'.

Besides newsfeeds which filter user content, social media also produce countless recommendations for users. Examples include which items to buy, movies to watch, which music to listen to, travel opportunities and who to invite into their social network. Hence, social media deploy extensive recommendation systems to match user preferences with products and services from a large number of candidates. According to Colace et al. (2015) recommendation mechanisms are composed of one or more of the following components. The first is pre-filtering, whereby subsets of items that are good candidates to be recommended are selected for each user. The second is ranking, whereby each item is ranked according to the predicted level of user preferences using well-known recommendation techniques such as content-based, collaborative filtering or hybrid algorithms. The third is post-filtering, whereby some items are dynamically excluded from the recommendation list according to user feedback or other contextual information. Coalce et al. (*ibid.*) propose a novel type of recommendation system which also incorporated users' opinions and item sentiment. Public opinion always drives choices. It is important to find out what people think. This is considered as the fundamental design aspect of modern recommendation systems, especially in the social media environment.

Beside recommendation systems, other matching algorithms are also utilized by social media. On the one hand, Arvidsson (2006) researches Match.com, a social media site for dating and

its matching algorithm called 'Venus' which automatically alerts users of others with compatible preferences. On the other hand, Ilten (2015) examines Sparked, a social media site for volunteering and its matching algorithm which alerts non-profit organizations when volunteers with the right skills they are looking to become available. These matching systems work in similar ways as they make certain objects visible to users, non-profit organizations in the case of Sparked, according to predefined criteria. Match.com entails characteristics of romantic relationships users are looking for and similar to Sparked this entails types of skills non-profit organizations require. Hence, volunteers do not bring their whole identity into the platform, only their skillsets and the ability to deliver products and services at specific times. Personal information is rarely exchanged. Users only disclose personal information about their professional affiliations, interests and enthusiasms but not about other aspects of their life.

Interface

Social media apply their algorithms to their databases to construct interfaces for user interaction. Kim and Mrotek (2016) analyze functionality and elements embedded into the interfaces of numerous online health communities (e.g. topic filtering systems, content moderation, user profiles, membership histories and interoperability with major social media platforms) and discover that best practices are rarely implemented. So how best should one select functionalities and elements to embed into social media interfaces? Numerous papers examine this issue. Rose and Oaebo (2010) suggest that it depends on the purpose of social media. Predefined topic categories may steer discussions in specific directions, whereas dynamically developed categories can increase flexibility. Synchronous debates (e.g. chat rooms) encourage short discussions while asynchronous systems can host more reflective and well-argued debates as participants have more time to think through their arguments. Lastly, while strict identity control increases entry course, it may also increase deliberation quality

and complete anonymity may encourage extremist and hate speeches. Spagnoletti et al. (2015) argue that social media design depends on whether the priority is for information sharing, collaboration or collective action. For example, social media for information sharing needs to be interoperable with major social media such as Facebook and Twitter to encourage circulation of news and update, while social media for collective action should provide a safe and secure environment for users to exchange information and reach consensus. Papachrissi (2009) points out that social media is designed to influence the behaviors of users. For example, taste ethos of LinkedIn is professional as it provides templates of self-presentation that follow resume formats for users to fill, while Facebook is more playful, providing various props and applications for users to construct their identities such as quizzes which allow friends to compare likes and dislikes.

Importantly, interfaces of social media are organized with social data. On the one hand, two authors conducted experiments on users of MovieLens, a web-based movie recommendation site where members rate movies and write movie reviews and recommendations. One problem with MovieLens is that over 20% of the movies listed have scant ratings and recommendation algorithms cannot make accurate predictions as to whether subscribers will like them. Ling et al. (2005) point out that social data can be used to encourage contributions by making users feel that they are unique; weekly messages can be sent to each user proclaiming their individuality by highlighting movies they rated favorably which few others liked. Ren et al. (2012) suggest that an increase in identity-based attachment to a group within an online community or bond-based attachment to an individual member of the community would increase attachment to the large community as a whole which, in turn, would increase member participation and retention. Users of MovieLens are assigned into different groups. Social data derived from top movies rated by different groups with low ratings from other groups and numbers of new ratings in the past week by each group are

provided. This information is intended to foster identity-based attachment and competition between groups and rating agreement and disagreement between users fosters bond-based attachments.

On the other hand, Butler et al. (2014) opine that lower participation costs and higher topic consistency cues can increase community size and resilience. Social data can be used to reduce participation cost and heighten topic consistency cues in numerous ways. For example, they note that allowing messages to be sorted by number of replies or recency decreases participation cost while revealing the number of replies also increases topic consistency cues as it helps users to gauge what topics the community finds interesting.

Networks have also been extensively studied as organization mechanisms for users and contents on social media. Tim Berners-Lee applied hyperlinks to the growing collection of documents on vast computer networks and the internet was born whereby documents are organized as a network. Zimmer (2009) asserts that this freed the reader from the 'straitjacket' of fixed and hierarchical systems of information organization, allowing more open-ended and non-deterministic data navigation. Other authors argue that this kind of information organization (i.e. network) is not neutral and cast implications on user behavior, while Sundararajan et al. (2013) suggest that networks on the internet have a particular kind of enduring structure which follows power law and they are often clustered with some nodes obtaining better positions than others in terms of node centrality.

Furthermore, two types of network, social and product have been examined through social media. Online and offline social networks differ. According to Haythornthwaite (2002), latent ties are created across each and individual pairs of users in social networks are maintained through social media. The potential is huge given the size of popular social media sites such as Facebook. This latent tie can be converted into a weak tie, which, in turn, can potentially be turned into a strong tie. Online social network, once created, shape the

kind of information users consume. Wohn and Howe (2016) assess the importance of online social network member diversity (age, race, nationality, occupation etc.). Users who do not have a particular diversity in their online social network are likely to be unaware of issues related to that diversity. For example, people with no ethnic diversity are more likely to be unaware of certain ethnic issues. However, awareness is not related to attitudes. If a user already has an opinion on a topic, he/she will be unlikely to change this. This finding concurs with Porter and Hellstein (2014) who examine chains of comments between users. Online debates rarely end with changes in opinion; rather, the positions of each user are solidified through elaboration of their positions via debates. The oldest example of a product network is the co-purchase network of Amazon ('Customers who bought this item also bought …') which makes product complementarity relationships explicitly visible. Oestreicher-Singer and Sundararajan (2012a) determine that categories of books that are highly and evenly influenced by product networks show consistently strong demand and revenue.

But why do users who participate in social media produce increasing amounts of social data? Skageby (2009, 2010) explain this by viewing social data as gifts. Receiving comments or likes is similar to receiving gifts from friends and these are reciprocated. Generalized reciprocity is prevalent in social media where comments and likes are given as gifts without an explicit agreement on the nature, value or timing on the return of the gifts. Exchanging comments and likes strengthens the social bond between users which is converted into social capital. Ellerbrok (2010) discusses numerous benefits of social capital. Importantly, it generates access to information and resources whereby small scale activities can be mobilized. For example, a young woman may use her Facebook page to let her online community know about her yoga class, generating an economic windfall. Interestingly, empowerment in this dimension comes with exploitation in other dimensions as social data

are being used for marketing by advertisers and surveillance by government. Ellerbrok (*ibid.*) defines empowerment as "the precondition for the collection and sale of unprecedented amounts of intimate data" and "a carefully orchestrated system; designed to encourage the willing and comfortable revelation of day-to-day intimacies over an extended period of time, explicitly for financial benefit of major institutions". Both businesses and political parties advertise themselves through social media. Boerman and Kruikeimeter (2016) and Kruikemeier et al. (2016) study user response to promoted tweets sent by brands and political parties. They find that users engage with content promoted by political parties less if they notice the heuristic cues used to promote the content, for example, 'Promoted by'. This is because the realization triggered persuasion knowledge and reduced the trustworthiness of the senders. As for businesses, persuasion knowledge is likely to be triggered anyway regardless of whether the heuristic cues are noticed.

Besides using social media for promoting contents, hoping that users do not notice the heuristic cues, political parties also employ social media in other ways. For example, 'target sharing' in 2012 whereby over 600,000 Facebook friends of the Obama campaign signed up for an Obama for America application that allowed automatic sharing of content from the Obama campaign with their friends (Bennett, 2015). Users are more likely to engage with shared content if it comes from their friends and persuasion knowledge is not activated. Hockman (2014) notes out that information streaming is the core of social media as a "dynamic, continuous flow of items that keeps updating according to new data that arrives from multiple, time varying sources" leading to presentism whereby "present has become the most crucial ordering mechanism of contemporary society". In the same vein, Elmer (2012) points out that Facebook and Twitter tend to bury 10 minutes old communication. On Twitter, 92% of retweets occur within the first hour so once an hour has passed messages become 'ancient history'. Twitter has thus become a key site for rapid response to live

political events. For example, 'fact check' can be posted over the course of live debates in real time with periodic links to more extensive information posted on party websites.

Other aspects of social media advertisement have also been researched. On the one hand, Kim et al. (2016) examine advertisements on the Newsfeed and Timeline of Facebook and determine that users view desirability focused messages more favorably in Newsfeed but feasibility focused messages more favorably in Timeline because psychological distances between users and messages are likely to be greater in Newsfeed than in Timeline which is a relatively private space accessible mainly by close friends and family. On the other hand, Buchanan (2015) investigates whether advertisements with violent content on Facebook induce higher levels of aggression-related cognition compared to non-violent advertisements. Results indicate that exposure to advertisements with violent contents increase the number of aggression-related words. However, this finding has no relation to whether violent advertisements lead to actual aggression.

Another related area of study is information adoption on social media, as the extent to which people accept content that they are presented with as meaningful after assessing its validity. On the one hand, according to Zhang and Watts (2008), this depends on argument quality and source credibility. In turn, this relationship is moderated by information retrieval systems. If the retrieval system is effective at selecting a limited set of relevant results which users can process, then the impact of argument quality on information adoption will increase relative to source credibility. On the other hand, Stavrositu and Kim (2014) study the impact of social media metrics (e.g. number of likes, shares and comments on Facebook, Google+ and Twitter) on adoption of health information. They assume the existence of a common misconception that individuals view themselves as better than average (i.e. third person perception). If users have this perception, they do not adjust their behaviors according to health information they receive because they regard any health threat (e.g. cancer) as not

applicable. A high level of social metrics helps to persuade users to adopt health related recommendations because this eases the third person perception, creating a bandwagon effect that decreases the perceived distance between users. In a similar vein, Lee-Won et al. (2016) determine that social media metrics play an important role in shaping injunctive norms or personal belief regarding what is approved by others and what ought to be done. However, research in this area is not conclusive. For example, Winter et al. (2015) demonstrate that numbers of likes do not influence how readers evaluate news stories.

Many authors have studied social media metrics and much ink has been spilled on metrics (likes) and social support. On the one hand, Carr et al. (2016a) and Carr et al. (2016b) describe how metrics can be perceived as providing the amount of social support that users are seeking. However, perceived social support becomes lower with higher perceived automaticity, whereby there is little rational or cognitive thought behind the provision of likes. Furthermore, perceived social support differs between sites, for example, higher metrics on Twitter are considered as least effective in delivering social support as compared to Instagram, Pinterest and Facebook. On the other hand, Wohn et al. (2016) unexpectedly find that users with high self-esteem have strong perceived social support from social media metrics. This represents 'rich get richer' dynamics, whereby people who already feel confident with themselves are more likely to perceive social support via higher social media metrics. More in-depth interactions may be required for users with low self-esteem to feel supported.

Besides likes, authors also explored the implication of other kinds of social media metrics. Tong et al. (2008) study the number of friends and social attractiveness whereby too much connection may result in negative judgment and cite Donath and boyd (2004) who coin the term 'Friendster whores' or reaction from people who realize that they are being invited to join someone's network of friends because they just provide an addition to a collection of

links. Chen et al. (2011) examine Karma points and moderation on Slashdot. Karma points measure a user's reputation based on the quality of their past comments. If users with low Karma points are being monitored more intensively than those with high Karma points, some opportunistic high Karma point users may perform worse than those with low Karma points, resulting in a reputation oscillation effect.

This reputation oscillation effect, among others, demonstrates how users game the system in Slashdot. Beside this, there are also other demonstrations on how users game the system. On the one hand, Crawford and Gillespie (2016) examine flags as tools for decentralizing content regulation tasks. Alas, users coordinate to manipulate the system. For example, a group of bloggers angered by the presence of pro-Muslim content on YouTube initiate a post called 'Operation Smackdown'. They orchestrate their supporters to flag specific pro-Muslim contents on YouTube by setting up playlists of YouTube targeted videos and celebrating the number of targeted videos that are removed. On the other hand, some users operate the EdgeRank algorithm of Facebook to ensure that they remain visible in the sites. Goodwin et al. (2016) point out that many young people share stories and photos depicting drinking and drunken behavior on social media sites because these attract more engagement. Carah and Dobson (2016) document the strategy of nightclubs to increase their visibility on social media by recruiting 'hot girls' to visit their clubs and circulate images of themselves and their peers on social media to boost engagement, leading to more attendance and alcohol consumption.

One of the ways nightclubs identify 'hot girls' is by searching for them by geographical locations. Indeed, social media has gone mobile. Goggin (2014) assesses Facebook's mobile features and discovers that it has become a locative form of mobile media with high reliance on location-based and mapping features. Facebook also claimed to have 65 million mobile users in 2009. Ghose et al. (2012) examine the implication of mobile interfaces compared to personal computers (PCs) regarding user behaviors. First, higher ranked links are more likely

to be clicked because of the effort required to scroll down a list of items. This can be interpreted as search cost which is higher for mobiles since they have smaller interfaces. They find a negative and statistically significant relationship between the rank of posts and clicks on them which is much stronger for mobile users than PC users. Second, mobile users tend to browse more geographically proximate brands than PC users because location-based services like "where's my nearest…" become more useful in a mobile context.

Social media also steps in to shape romance. Ben (2007) investigates the categorization employed by Gaydar, the largest UK gay dating website. He determines that users can describe themselves as 'bear', 'cub', 'builder', 'footballer' and so on but not 'camp' or 'effeminate'. Therefore, the site does not cater for members of the latter groups who would need to mislabel themselves to participate. MacKee (2016) examines usage of Tinder, a popular dating application by gays in London and finds the site to be a gay haven for connecting men looking for a genuine relationship. This stems from the organization of Tinder, whereby users login with a Facebook account to create an account on Tinder. Tinder then extracts user information from Facebook (e.g. image, name, likes and friends) and constructs user profiles. According to one informant, the Tinder profile is something that 'users can show to their mother'. This is unlike Grindr and most gay hook-up applications, where any picture can be uploaded to create hyper-sexualized spaces. David and Cambre (2016) examine the Tinder interface, whereby both users need to swipe each other profiles to enable matches and direct messaging. This entails quick thumb movements and sometimes users may make mistakes because of involuntary inflexes. Tinder capitalized on this by introducing Tinder Plus with a rewind feature, allowing the reversal of an undesired swipe. Lastly, Ridder (2015) discusses how youths used a social networking site in Belgium. He finds that 86% do not define their sexual identity, significantly more males (19.8%) revealed

their sexuality than females (8.8%) and more users looking for a relationship (25.6%) define their sexuality than users who are not (10.2%).

Research related to social media interfaces has also been conducted for various non-private domains. First is the domain of third-party developers. Social media needs to maintain interfaces with both people and machines. On the one hand, Bucher (2013) examines the importance of APIs for Twitter, launched just two months after its establishment in July 2006. APIs allow Twitter to be on every mobile platform and its search engine, initially built on APIs, was originally called Summize and acquired by Twitter in 2008. Twitter has maintained a vibrant ecosystem of third-party developers. According to ProgrammableWeb, a staggering 75% of Twitter's traffic came from third-party applications in 2010. On the other hand, Claussen et al. (2013) discuss the problems that Facebook faces from a flood of low-quality third-party applications. Facebook implemented rule changes in February 2008, whereby the amount of notifications that applications can send is determined by how frequently these messages are clicked on, a useful proxy of an application's ability to attract and retain users. After the rule changes, user ratings of Facebook applications improved, demonstrating the effectiveness of 'soft' quality incentives.

Second is the domain of enterprises. An enterprise social network operates just like an ordinary social network but is only for employees of particular companies. Users can add contents and metadata to those contents (e.g. tags, likes, comments). Enterprise social networks organize newsfeeds for each user according to the metadata. Users can also be alerted when changes are made to conversations in which they are participating (Majchrzak et al., 2013). Leonardi (2014), Leonardi (2015) and Fulk and Yuan (2013) discuss the implications of this on organizations. Importantly, it makes conversations invisible. Employees who use enterprise social networks tend to maintain connections with coworkers whom they do not know or might not regularly interact with offline. Discussions with the

coworkers appear in their newsfeeds. This enables employees to become more aware of 'who knows what' and 'who knows whom'. In turn, this facilitates at least two changes. Firstly, it helps to avoid duplication and encourages more effective reuse of knowledge and secondly, it enables employees to recombine ideas culled from various coworkers into new ideas to better solve problems.

Third is the domain of politics. Salender and Jarvenpaa (2016) document the constraint placed on Amnesty International by social media. As Amnesty International established a social media profile, the number of its digital supporters increased. Then, it launched digital petitions whereby supporters are solicited to sign and share petitions with their social networks. Because of the limited space for messages posted on social media sites like Facebook and Twitter, the accuracy of petitions can be compromised and comments can be added which are not necessarily in line with Amnesty International. However, much progress has been made regarding social movements (Zappavigna, 2011; Cardullo, 2015; Haciyakupogiu and Zhang, 2015; Milan, 2015; Yang, 2016). The social importance of the Twitter hashtag has been researched in detail and clicking on a hashtagged word enables retrieval of all messages containing that specific hashtagged word. This enables collective sense-making as focal themes rise to the surface among the whole population of chaotic tweets. For example, hashtags being used during the Egyptian revolution included #Jan25, #Tahrir, #Mubarak and #Egypt which demonstrated the desire of the Egyptian people to gather at #Tahrir Square on #Jan25 to dispel the dictatorial #Mubarak regime from #Egypt (Oh et al., 2015).

Reaction against commercial social media

The previous section discusses commercial social media such as Facebook, Twitter and Google+ which have been lambasted in numerous ways. Langlois (2015) stresses the need for critical research on social media which is lacking as most investigations are either

commercial (e.g. Facebook's notorious mood manipulation experiment) or state-sponsored projects to develop surveillance tools. Kennedy and Moss (2015) point out that data power needs to be democratized which can be done in three ways. First, the data mining tools that social media uses must be subjected to greater public supervision and regulation. Second, technologies of data mining (software, expertise and data) should be made available and accessible to the public. Third, data mining should be used in ways that enable members of the public to understand each other and consider issues that matter, fostering more reflective and active agents. These initiatives would entail a shift whereby the public is subjected to data mining practices of social media allowing more active and reflexive public agents.

Gehl (2015) suggests that an alternative to commercial social media already exists. So-called alternative social media have burgeoned in the last five years with sites such as Lorea, GNU social and Diaspora appearing on federated servers across the internet. ID2NT, Galaxy 2 and Visibility have appeared on the Dark Web which is a network only accessible via special software such as Tor. Peer-to-peer microblogs such as Twister and SOUP have been developed and installed on phones and computers around the world. These alternative social media share at least three characteristics situating them at the opposite end of the spectrum to commercial social media. First, there is no advertising. Money does not equate with influence in the alternative social media which deny the entire technical and organizational infrastructure that underpins online behavioral advertising. Second, alternative social media allow users to access more than just interfaces and guide users beyond filling in profiles, sharing/liking contents and friending to practices such as coding, administering and organizing the very system that enables these interface-level activities. Access to the underlying technologies is possible because alternative social media are built with free or open source software. Therefore, users can modify the program to suit their own requirements and even share copies with friends. Third, alternative social media deny

surveillance. They allow more playful identity construction as compared to commercial social media like Facebook, which is notable for its longstanding emphasis on real identities and social connections of importance to marketers. Alternative social media do not enforce real world identities and allow for experimentation on pseudonyms. Arguably, this allows users to take off their daily masks and becomes closer to who they really are.

Conclusions and future research

As everyday life becomes increasingly infiltrated with social media, the number of research papers on the subject has increased. While the majority of authors concentrated on the consumption side of social media (why, how and the consequences of the use of social media), research on the production side or operational aspects is still limited. Yet, this is a small, but growing portion of social media research. While research on the consumption side starts to gain traction by 2007 and further accelerates in 2013, research on the operational aspects grows steadily from 2013. It is imperative to understand the operational methodology of social media, whereby interfaces are assembled by the underlying algorithms and social data with the method of assembly casting real implications upon society and the behavior of users as a whole.

More research is required on all three fronts which constitute the operational aspects of social media. Black box algorithms as newsfeed interfaces can be opened and their implications assessed through investigating the source codes of alternative social media platforms. This can be done because alternative social media platforms are built with free or open source software inviting researchers to look beyond the interface. This cannot be done for typical, commercial social media due to its proprietary nature. On the data front, researchers should look beyond typical social data as documented and discussed by Kallinikos and associates (Kallinikos and Constantiou, 2015a; Kallinikos and Constantiou, 2015b; Alaimo and

Kallinikos, 2016). Data are also automatically input for users by, for example, numerous biometric sensors attached to their bodies. How do these sensor data differ from the usual social data whereby social interaction is stylized and what are the implications of utilizing this kind of data? Lastly, one can apply the methodology employed to study the online communities to typical social media like Facebook. Kim and Mrotek (2016) analyze the functionality and elements embedded into the interfaces of different online communities. What is the functionality of the elements embedded in Facebook's interface or other typical social media, how are they being constructed and what are their implications for user behaviour?

Ranking and Information Display in Social Media Platforms

Akarapat Charoenpanich, LSE

Abstract

This article applies the theory of digital objects as postulated by Kallinikos et al. (2013) to elucidate the information display techniques applied on social media platforms. It posits that social media information displays are assembled using two components as databases and algorithms. Algorithms operate on the social data collected to assemble information displays for user interaction. These, in turn, generate even more feedback into the social media database. Three theoretical claims regarding digital information displays have emerged. First, ranking is recognized as the dominant strategy to display information on social media platforms. Second, navigation through the information display is an intensely interactive experience. Third, the information display is dynamic, fluid and unstable. These have culminated as the interaction system discussed by Wegner (1997). The information display strategy of Last.fm, a popular social media platform for music discovery is then presented and analyzed to assess these three theoretical claims. The first two theoretical claims are confirmed by data from Last.fm, while the assessment of the last claim is surprising as the information display assembled by Last.fm is proved more stable than initially expected. However, this stability does not last as a digital information display is, after all, a digital artifact which is always in constant flux. These dynamics lead to interesting implications for user choice, behavior and demand for music on Last.fm.

Introduction

Social media has become increasingly important in our everyday lives. In June 2017, Facebook claimed two billion active users, representing 27% of the global population. Social media platforms are emerging as central facilities through which people interact and search for content on the internet. This is why it is important to study how social media platforms organize information and construct information displays for their users as this can influence user behaviors, leading to huge societal consequences. This study theorizes and analyses information displays as constructed by social media platforms which is a timely and important study topic.

This article is divided into two sections. The first is theoretical and presents a conceptual framework of the information organization of social media platforms. While much ink has been spilled studying the consumption side of social media (i.e. why people use social media, how people use social media and the consequences of using social media), study of the production side or the inner operational aspects of social media is still relatively limited. However, the literature on the operational aspects of social media platforms is growing, indicating that this is a fruitful area of study. This article examines the operational aspects of social media platforms and asks two main research questions: *(1) How do social media platforms construct their information displays?* and (2) *What are the characteristics and consequences of information displays constructed by social media platforms?* The theory of digital objects postulated by Kallinikos et al. (2013) is used to scrutinize the information displays of social media platforms. This theory posits that information displays are assembled by two basic components as a database which stores social data and an algorithm. A heuristic conceptual framework is presented to scrutinize the inner operations of social media platforms which does not only take data (e.g. Alaimo and Kallinikos, 2016), algorithms (e.g. Bucher, 2012) and interfaces (e.g. Ren et al., 2012) into account but examines all three

components together to ascertain how social data is combined with algorithms to construct the information displays of social media platforms.

Three theoretical claims have emerged regarding digital information displays. First, ranking is recognized as the dominant strategy to display information on social media platforms. Second, navigation through the information display is a highly interactive experience. Third, the information display is dynamic, fluid and unstable. Interactivity of social media platforms is very important in the digital information environment. Wegner (1997) points out that "interactive systems are more powerful problem-solving engines than algorithms". While the latter "yield outputs completely determined by their inputs", the former "provide history-dependent services over time that can learn from and adapt to experience". Interaction is irreducible to algorithm and it can be argued that social media platforms present users with interactive systems composed of seemingly limitless numbers of interfaces which can be activated by users to serve different information needs as they arise.

The second section assesses these theoretical claims against empirical data concerning information display strategies collected from Last.fm, a popular UK-based social media platform for music discovery. Results suggest that ranking (ordered list) is indeed a dominant strategy to display information on social media platforms. In addition, the diverse types of information display utilized by Last.fm demonstrate that digital information displays are highly interactive although they are all rankings. As noted, a digital information display should be more fluid and unstable than its physical counterpart. Therefore, it is also important to investigate the dynamics of these rankings. Therefore, artist similarity rankings were collected over time from Last.fm and analyzed. Arguably, this particular type of ranking is central to the operation of Last.fm as the input of personalized recommendations. Interestingly, these rankings seem to be much more stable than initially expected and they also stabilize over time during the business-as-usual operations of Last.fm. However, they

can destabilize as the business plan/goal of Last.fm changes and have important and interesting implications for user choices and behavior.

Figures 1 and 2 present the empirical findings which emerge from this research. Last.fm ubiquitously utilizes music charts (ranking) to display music. Figure 1 shows weekly charts (track version) based on aggregated listening data for all genres of music submitted in the UK for the week ending 5th July 2015. The top 'hype' track ('Freak of the Week' from 'Krept & Konan') achieved a 'hype score' of 3,140 (in the left column) and the most popular track ('Lean On' by 'Major Lazer') achieved the highest number of listeners at 1,785 (in the right column). Although Last.fm does not disclose the formula underlying these 'hype scores', this presumably has to do with the rate of growth of the number of listeners and indicates 'rising' artists and tracks. Charts can also be produced for different time periods, geographical locations and music genres. Thus, the digital information environment of Last.fm is highly interactive. Figure 2 scrutinizes the stability of artist similarity rankings. The top one thousand artists are gathered for November 2014 with artist similarity rankings analyzed over time. Figure 2 plots the number of artists in the rankings of similar artists in June 2015/December 2015 which appeared in rankings of similar artists in November 2014/June 2015. The chart demonstrates that artist similarity rankings were more stable between November 2014 and June 2015 than between June 2015 and December 2015. The second section of this paper delves deeper into these findings and provides an overview of other kinds of rankings assembled by Last.fm and the reasons behind the changes in stability of artist similarity rankings.

Figure 1: Music charts based on aggregated listening data

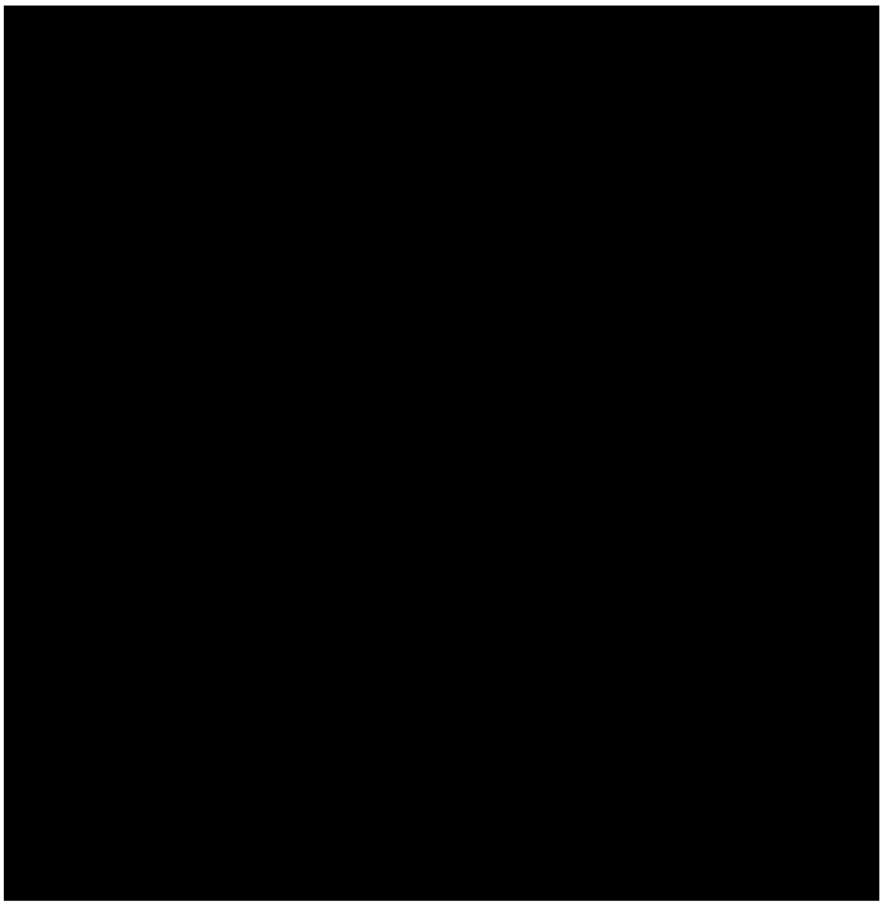
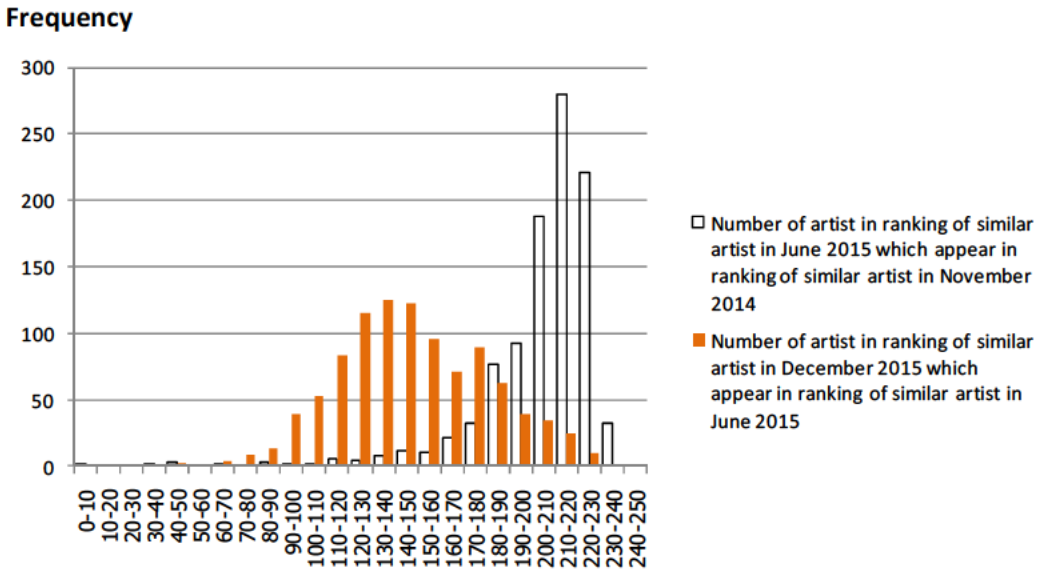


Figure 2: The dynamics of similarity rankings



A conceptual framework of information display on social media platforms

Social media information displays are assembled from databases and algorithms which can be regarded as digital objects in constant flux. The conceptual framework of this article begins with the theory of digital object as proposed by Kallinikos et al. (2013) before moving on to discuss the components which constitute the information displays of social media platforms, namely, the information interface itself and its underlying database and algorithm. Finally, three conjectures regarding social media information displays are evaluated using data from Last.fm, a social media site for music discovery.

The theory of digital objects

Kallinikos et al. (2013) propose that all digital objects share four generic attributes, making them less stable but more malleable compared to everyday physical objects. These four attributes are editability, interactivity, openness and distributedness which presuppose more fundamental constructions as modularity, whereby digital objects are composed of loosely-coupled components or modules and granularity, whereby any digital object can be decomposed into fundamental elementary binary units. Likewise, social media information displays are assembled from two basic components as databases and algorithms. These are both digital objects and in constant flux, thus information displays of social media platforms are also in constant flux. Databases can be continuously updated with new streams of information and algorithms can be adapted to optimize the business of social media. According to Kallinikos et al. (2013), digital objects are performative; therefore, their changes will also affect user interaction. As information displays become unstable, user behaviors are also likely to become unstable. The following section delves deep to discuss the

components which constitute the information displays of social media platforms. How do social media platforms collect data from their users? Which algorithms do social media platforms deploy to assemble information displays for user interaction and how do these changes cast implications on information displays?

Data

Social media collect data from two sources; firstly from users who interact directly with social media platforms. Alaimo and Kallinikos (2016) recognize three kinds of social data that social media sites collect from their users. First is profile data as the demographics of users including race, location, gender and age. Second is behavioral data, whereby user participation is shaped by the structural attributes of social media platforms. For example, social media platforms set up standardized user activities including tagging, following and liking. This, in turn, leaves computable and countable data footprints behind to be further processed and used for assembling different types of information display. Alaimo and Kallinikos (*ibid.*) regard behavioral data as the most valuable kind of 'discrete and granular' social data since its creation entailed drastic simplification compared to activities in everyday life outside social media platforms. Countable data can be very easily further processed. Facebook's 'like' feature is one example. Facebook regards every 'like' as being the same, but this feature may have a different meaning for people in actual everyday life. This is primarily because of its codification which entails a simple premise that everything (e.g. users, comments, photos) can be regarded as an object and every object can be connected by pre-programmed standardized activities such as tagging, liking and following. The third kind of social data is user-generated contents, for example, discussion of users within an online community. This data is often stored in a database and seldom used. Behavioral data is also the most important kind of social data because it makes profile data and user-generated

contents more useful. For example, user-generated content such as images become computable once they are mapped with behavioral data.

Secondly, social media platforms also source their data from the broader digital ecosystem. Helmond (2015) points out that social media sites are officially transformed into social media platforms when they establish their application programming interfaces (APIs), rendering the platforms reprogrammable by third party developers. In the mid-2000s, social media sites began to establish their own APIs. For example, Delicious established APIs in 2003, Flickr in 2004 and Last.fm, Facebook and Twitter in 2006. Developers can access platform data and functionality through APIs, enabling them to read, write and delete user data. There has also been dissemination of the so-called widgets as plugin modular components enabling integration of platform content and functionality into another website with a few lines of code. This includes social plugins such as the Like button developed by Facebook. Technically, these social buttons function as APIs by sending specific requests to Facebook's platform, for example, to record the number of people who like the post or to publish 'like' on user timelines once they click the Like button. According to Gerlitz and Helmond (2013), this has given rise to the 'like economy'. Facebook has extended beyond the limit of its platform by offering widgets which can amalgamate other websites and applications. In other words, Facebook, as a social media platform, has begun to source its data from the broader digital ecosystem.

Algorithms

Many algorithms cast significant implications on the society. These include PageRank and EdgeRank algorithms which were critically scrutinized by Introna and Nissenbaum (2000) and Bucher (2012). The recommendation algorithm is another; however, this has yet to be

closely examined. The empirical section of this article will attempt to examine the output from recommender systems. Therefore, before moving on, it is important to discuss recommender systems before the theory of information organization in the digital environment as postulated by Weinberger (2008) is considered. Arguably, the assembly of information display by the recommendation algorithm is one of the many kinds of information display which can be assembled in a digital environment.

Information retrieval, information filtering and recommender systems

Information filtering systems are a subset of information retrieval systems. Belkin and Croft (1992) suggest that information retrieval and information filtering systems can be viewed as different sides of the same coin. While an information retrieval system selects relevant documents according to one-off queries from users, an information filtering system selects relevant documents according to long-term user interest. In other words, a database of user profiles must be maintained for an information filtering system to operate. An information filtering system can be considered as a 'zero query' search of an information retrieval system. Many algorithms can match user profiles with documents to produce relevant data. According to Oard (1997), these include rule induction, instance-based learning, statistical classification, regression, neural networks and genetic algorithms.

There are three key classical information retrieval techniques as (1) exact match and Boolean model, (2) vector space model and (3) probabilistic model (Baeza-Yates and Ribeiro-Neto, 2011). Exact match and Boolean model are considered to be the weakest as they do not rank documents according to relevancy. Ranking documents is superior to presenting users with a set of documents, as ranking allows humans and machines to synergistically achieve better performance than either can achieve alone at enhancing user satisfaction (Oard, 1997; Winiwater et al., 1997).

Recommender systems can be considered as a subset of information filtering systems. However, many kinds of information filtering systems are not considered as recommender systems, for example, an email spam filtering system. More specifically, a recommender system employs instance-based learning to identify the relevant items to users, whereby the relevance of new items is assigned according to the relevancy of the most similar items (Pazzani and Billsus, 2007). Thus, the main tenet of recommendation algorithms is to 'recommend similar items to those that users already like'. To operate, a recommender system must construct a similarity between items to be recommended. This can be accomplished in two ways, giving rise to two categories of recommender systems as collaborative filtering recommender systems and content-based recommender systems. The former assign similarity according to correlation between ratings made by different users for different items. In other words, two items will be assigned with a high degree of similarity if they are being liked or disliked by the same group of people. The latter assign similarity according to the similarity of product features. Content-based recommender systems have been deployed mostly in the domain of text, where an algorithm exists to extract product features automatically (i.e. term frequency-inverse document frequency or TF-IDF). An algorithm to automatically extract meaning and features from sound, image and video has yet to be widely deployed.

It is important to note that a recommender system also ranks items. When a recommender system constructs similarity between the items to be recommended, this is similarity in terms of degree. Rankings of similar items exist for each and every item being recommended by a recommender system. This means that there are some items which are more similar to certain items than another item. Also, a recommender system assembles a ranking of recommended items to users ranked according to the level of predicted rating that users would have made if they were to consume the items. Arguably, this kind of display (i.e. ranking) is superior to

presenting users with a set of items as it allows humans and machines to synergistically achieve better performance and enhance user satisfaction (Oard 1997; Winiwater et al., 1997).

Theory of information organization by Weinberger (2008)

Weinberger (2008) discusses the theory of information organization by juxtaposing information organization in digital and physical environments. The main distinction is that information organization in the digital environment entails 'sort on the way out' strategy which is opposite to the physical environment. The amount of metadata which can be attached to items in the digital environment is seemingly unlimited. This differs from the physical environment whereby too much metadata would become incomprehensible, like swollen card catalogs for books in libraries. A librarian needs to carefully assign metadata to books and not all metadata thinkable by librarians will be recorded as some would need to be filtered out beforehand. In other words, a physical item can only be in one place (e.g. a book has only one place on a bookshelf), while digital items can be in many places. Weinberger (*ibid.*) suggests that a prototype classification replaces the Aristotelian classification in the digital environment whereby the boundaries of different objects become blurred. There is no longer a clear-cut boundary which separates one object from another.

Seemingly limitless amounts of metadata attached to items in the digital environment represent social data stored in databases on social media platforms. Algorithms can then operate on the data and 'sort on the way out'. If a recommendation algorithm is applied to the data, then the output becomes ranking of similar artists and ranking of recommended items. However, a recommendation algorithm is not necessarily applied and there are also other forms of algorithms. Grosser (2014) points out that social media platforms are often filled with metrics which can be defined as "enumeration of data categories or groups that are easily obtained via typical database operations and represent a measurement of that data".

Social media platforms often also use these relatively basic algorithms to assemble information displays. Facebook, for example, is filled with metrics such as numbers of likes, comments, shares, friends, mutual friends, pending notifications, events, friend requests, message waiting, chats waiting, photos, places and much more. Of course, Facebook produces more metrics than those revealed to users and what it chooses to reveal depends on whether such revelations would increase user engagement on its platforms.

As for more content-oriented social media platforms, it is common to see rankings of all sorts to display information. For example, many social media platforms use popularity ranking and its variations to display contents on their websites. This can be easily accomplished because rankings are computed from behavioral data which are already numerical in nature. Because there are seemingly limitless kinds of rankings assembled for users to browse through, information displays of social media platforms are highly interactive.

Information display

Although the number of imaginable visual representations of information displays is virtually infinite, Kleinmuntz and Schkade (1993) suggest out three fundamental characteristics that may cut across a broad range of displays. In principle, each of these characteristics may vary independently of one another. The first is the form of the individual information items. There are at least three different distinct forms as numerical, verbal and pictorial. Furthermore, there can be variations within given forms, for example, fractions, decimals, and scientific notation for numerical forms, single words, short phrases, everyday terminology and specialized terminology for verbal forms and charts (bars, lines, pies etc.), faces and other visual symbols for pictorial forms. The second is the organization of display items into meaningful groups or structures. One such organization is a table or matrix with

rows corresponding to alternatives and columns corresponding to attributes, whereby each entry can be in any suitable form (numerical, verbal and pictorial). One example of this type of organization is a consumer report and others are lists or paragraphs of text as travel guides and lists of hotels/resorts. There can also be more complex structures of organization display. One example is labels on consumer food products which may include both lists and tables. The third is the sequence (and ranking) of individual items or groups of items. A given organization of display items does not necessarily specify an order in which individual items or groups of items must appear. Lists can appear in many different sequences and rows or columns in a table/matrix can vary. There are many ways to put things in a specific order. For example, a common practice is to sort values in a bar graph so that bars appear in decreasing or increasing order. Moreover, information can be sorted into alphabetical or chronological order. Sequencing can be important as it often determines the order in which information is read by decision-makers; therefore, it may have implications for information processing. Kleinmuntz and Schkade (1993) view sequencing as also encompassing ranking; this appears strange to those who regard sequencing in a horizontal format, although this is merely a matter of terminology.

A body of empirical literature has studied the impact of numerous characteristics of information displays regarding the decision-making strategies of people (Schkade and Kleinmuntz, 1994). One stream studied the form of individual information items. Previous authors often compared quantitative and qualitative presentations of information, whereby equivalent information is entered with either numbers or words (Huber 1980; Stone and Schkade, 1991). Another stream looked at the organization of display items. Here, comparisons were often made between simultaneous versus sequential presentations of information, whereby subjects are presented with pages of information which either (1) present the values of all alternatives for one attribute on each page or (2) present the values of

all attributes for one alternative on each page (Bettman and Kakkar, 1977; Bettman and Zins, 1979; Jarvenpaa, 1989). Unfortunately, before the nineties, relatively little attention was paid to sequencing (and hence ranking) as a form of information display.

Arguably, one major shift in information display literature has been the increasing attention paid to sequencing and ranking in recent years. Seminal ideas include Espeland and Sauder (2007) who note the increasing demand for accountability, transparency and efficiency in the past two decades, leading to an 'audit explosion' (Power, 1994) and the emergence of 'audit culture' (Strathern, 2000). This has led to the creation of numerous quantitative measures to evaluate the performance of individuals and organizations, including cost-benefit analyses for public investment projects, standardized tests for students and schools, performance measures for firms, assessments for universities and ranking for schools, firms and hospitals.

In the context of the increased interest in quantitative measures, Espeland and Sauder (2007) studied USNews law school rankings. Commensuration constitutes an important mechanism which shapes the sense-making (cognition) and thus the decision-making of consumers by essentially transforming the qualities of multiple entities into quantities that share a metric. Individual entities then appear to be comparable to one another. The process of commensuration entails far-reaching simplification of the diverse constitutions of the entities being ranked and reduces the amount of information that people need to deal with. Simplification is accomplished in two ways. Firstly, it makes a vast amount of information irrelevant. This essentially entails disregarding differences between entities which cannot be expressed metrically (e.g. qualitative and tacit knowledge). Secondly, it integrates the remaining information associated with different entities into the same format, a shared metric. As such, commensuration unites entities by establishing precise relationships between them, making heterogeneity among them appear less visible. This, in turn, establishes statistical relationships between entities and allows rankings to be computed. Consequently, it is

arguable that ranking and commensuration can significantly transform the sense-making processes of people. While the world is naturally replete with equivocality (Weick, 1979; Weinberger, 2008), anything can therefore be unique, but anything can also become comparable with commensuration.

In general, being higher up the rankings is better for those who are being ranked. Espeland and Sauder (2007) proclaim that the USN law school rankings exhibit self-fulfilling prophecies. For example, student decisions are correlated with ranking, and these also determine the quality of applications schools receive. In addition, budgets can become tightly linked to ranking as departments compete for funds from other departments within the same university. Furthermore, being lowly-ranked makes it more difficult to solicit alumni, making it even more difficult to generate revenues and resources to bump up ranking. Espeland and Sauder (*ibid.*) go further to identify numerous ways in which law schools react to ranking. For example, they may funnel large amounts of money into activities that receive little funding to influence ranking criteria, such as merit scholarships and marketing. Espeland and Sauder (*ibid.*) also document a wide range of gaming strategies which entailed 'manipulating rules and numbers in ways that are unconnected to, or even undermine, the motivation behind them', such as reporting inconsistent sets of numbers to different agencies.

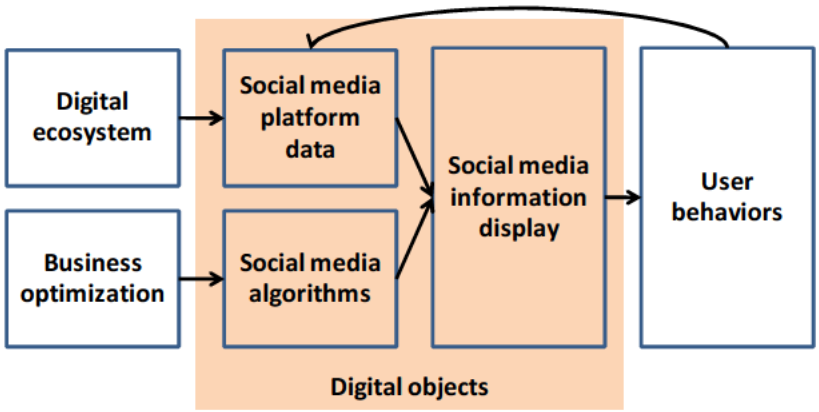
It is important to note that a broader definition of ranking than that provided by Espeland and Sauder (*ibid.*) is adopted here. Espeland and Sauder (*ibid.*) rank law school performance, which naturally contains a normative aspect, assuming that the positions of the social units being ranked capture differences in quality (e.g. better or worse law schools). However, it is often the case that information displays on the internet lack this quality. Arguably, they are merely ordered lists without normative judgment. Nonetheless, being highly ranked on ordered lists is beneficial in terms of visibility, though inherently neither better nor worse.

Here, a broader definition of ranking is adopted whereby ordered lists are seen as an instance of the wider category of ranking.

Theoretical synthesis and conjectures

Figure 3 synthesizes the discussions so far regarding social media information displays as digital objects and assembled components as databases and algorithms. Social media sites collect data from two sources. One is directly from user participation on its platform and another is through the wider digital ecosystem. Databases are unstable as they are being continuously updated with new streams of data. Algorithms operate on the databases to assemble information displays and they can also be unstable as they are tuned to optimize the business of social media. Instability of both databases and algorithms would then render information displays unstable. This, in turn, casts implications on user behaviors which are being captured as social data and fed back into the databases of social media.

Figure 3: Conceptual framework of information displays of social media platforms



Three theoretical conjectures then emerge from this conceptual model. First, ranking is supposedly the dominant strategy for social media platforms to assemble information displays. Recommender systems produce two main outputs. One is a similarity network and the other is a recommended list of items. Both can be represented as a ranking with similarity in terms of degree and recommended items ordered according to predicted levels of

preference. Arguably, this kind of display (i.e. ranking) is superior to presenting users with a set of items as it allows humans and machine to synergistically achieve better performance, enhancing user satisfaction (Oard 1997; Winiwater et al., 1997). There are also other kinds of displays which can be assembled by social media platforms through the utilization of simple computational procedures such as summation. Ranking emerges from this as these displays are assembled from behavioral data which are countable in nature. Therefore, arguably, social media platforms produce a seemingly limitless number of rankings for users to browse through. This leads to the second theoretical conjecture that experience within the digital information environment is extremely interactive. The third theoretical conjecture is that digital information displays are highly fluid and unstable. Again, this is because an information display is a digital object, which is naturally in constant flux and assembled by a database and an algorithm which are also digital objects. Because databases and algorithms are potentially in constant flux, the digital information display in which they assemble ought also to be highly fluid and unstable. While a database of social media platforms is constantly updated with new streams of social data, an algorithm is characterized by openness, and, so, can always be altered.

This has culminated in an interactive system for content discovery as discussed by Wegner (1997). Recommendation algorithms like any other algorithms are 'dumb and blind'. Because the main tenet of recommendation algorithms is to 'recommend similar items to those that users already like,' they can hardly be novel. Fortunately, recommender systems are not alone within the information environment maintained by social media platforms which also assemble rankings of all sorts. One ranking may be better at serving a particular user information need than another. Given that there exists a seemingly endless number of rankings which are assembled by social media, the information environment of social media platforms is likely to be able to accommodate the information needs of multiple stakeholders

over time. This is a truly interactive system, whereby users may activate particular rankings of their interests to make novel discoveries. As users do that, recommendation algorithms also become smarter. Wegner (*ibid.*) points out that 'smartness in mechanical devices is often realized through interactions that enhance dumb algorithms so they become smart agents'. As users make novel discoveries by unleashing the power of digital information organization as discussed by Weinberger (2008), novel discoveries are updated in user profiles. As recommender systems recommend items similar to those novel discoveries, the recommended items are also likely to be relatively novel, culminating in smarter recommender systems.

How true are these theoretical claims? In the next section, the information display strategy of Last.fm, a popular UK-based social media platform for music discovery is presented and analyzed to assess these theoretical claims. What emerges is that ranking is, indeed, the dominant strategy which social media platforms use to display information. They are also, indeed, highly interactive. However, interestingly, some of these rankings are found to be much more stable than initially expected. All these findings have important implications for user choices and behavior.

The empirical analysis of Last.fm

Last.fm, founded in 2002, is, now one of the oldest and most popular social media platforms for music discovery. Last.fm attempts to keep track of everything users listen to on the internet as its application programming interfaces (APIs) allow users to submit their listening data (behavioral data) from over 600 playback applications, services and devices. Essentially, Last.fm uses this behavioral data to assemble an information display for its users. The information display strategies deployed by Last.fm are presented and analyzed to assess the aforementioned theoretical claim. This empirical section is divided into two parts. The first part presents a static version of the overall information displays assembled by Last.fm.

What emerges is that they all are rankings. Navigation through these rankings is a highly interactive experience as there are many diverse types of rankings for Last.fm users to browse through to discover music. The second part traces the dynamics (changes over time) of one particular type of ranking assembled by Last.fm as the artist similarity ranking. Arguably, this particular type of ranking is central to the operation of Last.fm as it is also an input into the assembly of personalized recommendations. Interestingly, these rankings are found to be much more stable than initially expected and also stabilize over time during the business-as-usual operations of Last.fm. However, they can destabilize as the business plan/goal of Last.fm changes. These findings have important and interesting implications for user choices and behavior. Limited research has assessed the dynamics of information displays. For example, Dou et al. (2010) and Oestreicher-Singer and Sundararajan (2012) treat ranking in the digital realm as being static. This paper assesses the third theoretical claim and also fills this research gap.

An overview of the information display strategies of Last.fm

Thanks to database technology, Last.fm assembles a very diverse set of music rankings for their users to browse through, culminating in an interaction system as discussed by Wegner (1997). For example, Last.fm has ubiquitously utilized music charts to display music. One set of these charts is calculated based on listening data submitted by users. The charts are updated weekly and can be filtered by the geographical locations from which listening data is submitted (according to IP address) and by genres of music (according to tags applied to music by users). These charts list the most popular artists and most popular tracks, whereby artists and tracks are ranked according to the number of weekly listeners and 'hype' artists and tracks, whereby they are ranked according to 'hype scores' (Figure 1).

Besides these, there are also other types of charts utilized by Last.fm. One is the personal music chart of each individual user. Indeed, Last.fm enables users to store their listening

histories in one place and these are updated instantaneously as users submit listening data. Figure 4 shows the music charts of an individual user on last.fm. One chart ranks the artists and another chart ranks the tracks according to the listening data of the user. Anyone who visits the profile page of the user can view these charts to learn about his/her music taste. Therefore, it is possible to discover new music by browsing through the personal music charts of friends or strangers who leave comments or write reviews on Last.fm. Another is group music charts compiled by the listening data of all users in the group. On Last.fm, anyone can set up a new group and any other user can choose a group that they would like to join. Last.fm also compiles music charts for a group using the listening data of all users who join the group. Figure 5 ranks artists according to weekly listening data submitted to Last.fm from users in a group. Note that users may also select to view artists 'unique to the group'. This option ranks artists according to a certain score that measures the uniqueness of artists listened to by users within the group as compared to a general user of Last.fm. Unfortunately, the formula underpinning the score is not disclosed.

Figure 4: Personal music charts of an individual user

Figure 5: Group music chart

Figure 6 ranks artists according to a 'similarity score' or the degree of similarity between them and 'Coldplay' as constructed by Last.fm. Here, 'Coldplay' is selected to show how Last.fm uses a 'similarity score' to display artists. 'Keane', 'OneRepublic', 'Snow Patrol' and 'Imagine Dragons' are 'super similar' to 'Coldplay'. Although the formula underlying the calculation of a 'similarity score' is not disclosed, we know that tagging data is used alongside listening data as inputs into the calculation. Here, artists are construed as being more 'similar' if they are being listened to by the same group of users and tagged by the same label. For example, if two artists are overly tagged with the word 'Jazz', then they would be construed as being 'similar' by Last.fm. A 'similarity score' for the two artists will increase even further if they are also being listened to by the same group of users. Using only

listening data to compute a 'similarity score' is deemed insufficient for Last.fm. For example, classical music listeners can also be listeners of hard rock music. Consequently, hard rock music will be deemed 'similar' to classical music if listening data is the sole input into the calculation of a 'similarity score'. The addition of a tag as another input into the calculation of a 'similarity score' helps to alleviate this problem. Normally, the ranking of similar artists will be updated with new listening and tagging data every few months.

Figure 6: Similar artists to Coldplay

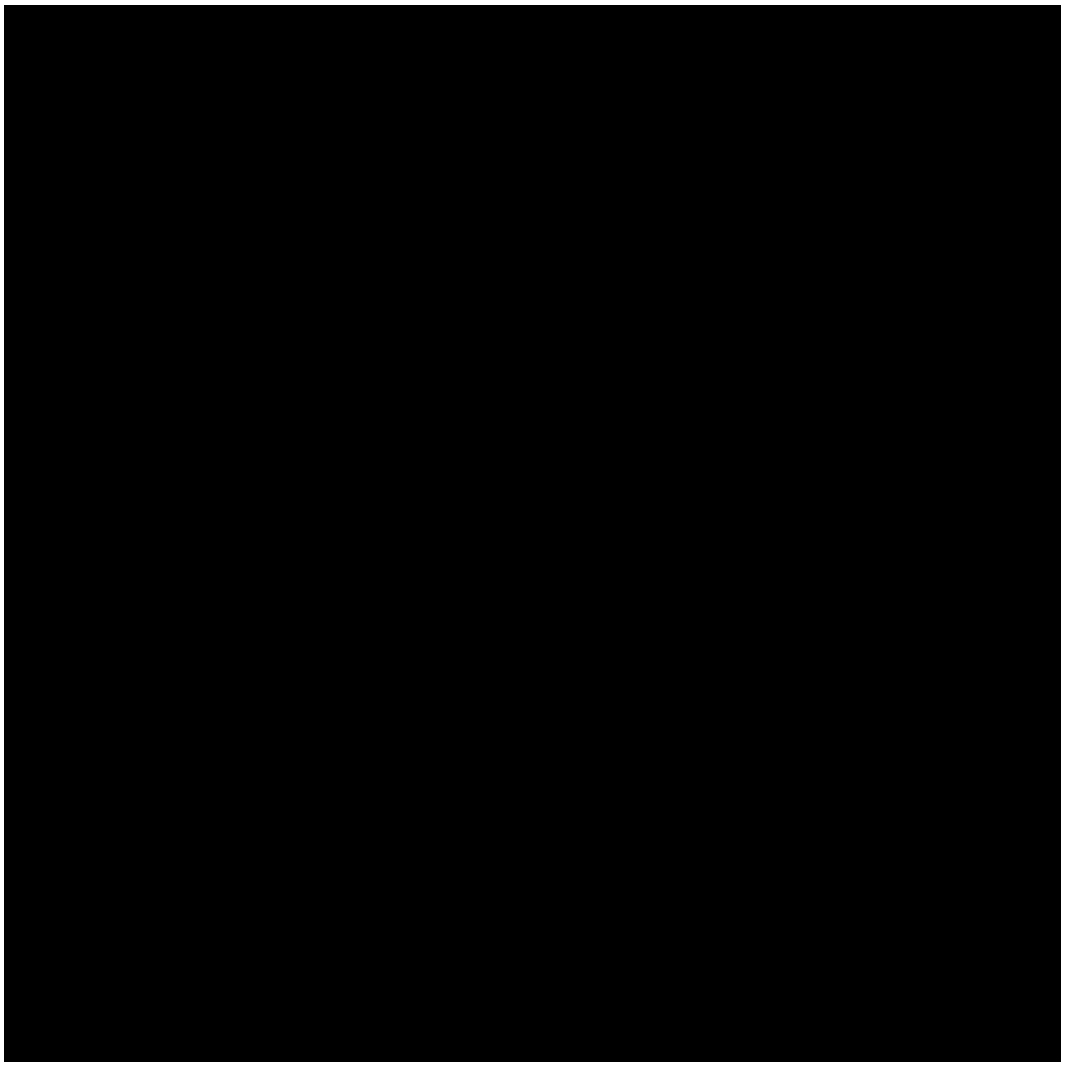


Figure 7 reveals my personalized recommendation page on Last.fm (as I cannot view the personalized recommendations of other users) On its website, Last.fm recommends as many as 90 different artists (on 6 pages) for me, as those it thinks I would like listed first. Indeed,

this is a ranking of suggested artists that is personalized for me. I have 'ETC.', 'Booker Ervin', and 'Paul Chambers' in the top three positions. All of these 90 artist recommendations share two characteristics. Firstly, I must not have listened to them more than a few times in accordance with my listening history. Therefore, they are likely to be new to me. Secondly, the recommended artists must be associated with similar artists who appear in my listening history. 'ETC.' appears at the top of my personalized music recommendation set. I have listened to the band only 3 times and they are similar to 'Crescendo', 'Friday', 'Lipta', 'Tattoo Colour' and 'Sqweez Animal' that I have listened to 58, 6, 7, 87 and once respectively. 'Booker Ervin' is in the second position on the ranking of my suggested artists. He is similar to 'Blue Mitchell', 'Jackie McLean', 'Sonny Clark', 'Tina Brooks' and 'Oliver Nelson' that I have listened 5, 5, 5, 4 and 4 times respectively.

This demonstrates how recommendation algorithms become smarter for content discovery as they are incorporated into the interaction system as discussed by Wegner (1997). 'ETC' can be considered as a safe recommendation for me. It is not very novel because it is similar to artists that I already listen to a lot such as 'Crescendo' and 'Sqweez Animal'. However, the recommendation of 'Booker Ervin' is relatively novel for me. This is because it is similar to artists that I listen to only a few times. I discovered these artists not through recommendation algorithms but by browsing other music charts assembled by Last.fm.

Although the formula underlying personalized suggested artists is not made public, we know that Last.fm utilizes a version of an item-item collaborative filtering recommender system to assemble personalized recommendations. Here, artists who are similar (according to artist similarity rankings) to other artists that a user likes (as gauged by number of listening data), will be recommended to the user. Henceforth, arguably, the artist similarity rankings are central to the operation of Last.fm as it is also an input into the assembly of personalized recommendations.

Figure 7: Artist recommendation page for the author

Dynamics of digital information display: a case of artist similarity ranking on Last.fm

To quantify the dynamics of the ranking of similar artists, it is possible simply to count the number of artists in each ranking of similar artists over a certain period who also appear on corresponding rankings of similar artists over the previous period. Before we can gather a ranking of similar artists to describe the dynamics, a sample of artists is required. Here, our sample of artists is the top one thousand artists as the rankings of most popular artists in November 2014. This is the longest length of rankings retrievable with the APIs of Last.fm. In other words, we use the rankings to assemble our artist sample. Some may contest that this is a strong selection, but the contrary may also be valid. Over approximately the same period, a set of representative samples of around twelve thousands users was gathered together with their listening data. Strikingly, I found that almost 40% of the listening data of the representative sample of users was associated with the artist sample in question, i.e., the

top one thousand artists in the ranking of most popular artists in November 2014. Therefore, this selection of samples was not too narrow: it correctly captured the artists that Last.fm users listen to.

Rankings of similar artists are updated every few months. Here, the rankings for the sample were retrieved from November 2014, June 2015 and again from December 2015 (three data points). Note that significant changes to the operation of Last.fm (and also presumably to the recommendation algorithm) occurred in September 2015. The dynamics of the rankings of similar artists between June 2015 and December 2015 inevitably reflected these changes. However, no such significant changes to Last.fm occurred between November 2014 and June 2015. Therefore, it can be argued that changes to the rankings of similar artists between November 2014 and June 2015 reflect the business-as-usual scenario of Last.fm. In this period, the database of Last.fm was merely updated by a new stream of social data with the algorithm underlying the recommendation algorithm remaining unchanged throughout the whole period. In addition, note that the rankings of popular artists differ from the rankings of similar artists in one important aspect. The former seem to be simpler and more universal. Although both rankings are ordered by quantitative data, the rankings of popular artists are assembled based merely on the number of listeners, while the rankings of similar artists are assembled based on 'similarity scores' which entail more complex computation. One ranking of similar artists for each individual artist exists and this is made possible only because of the database which underlies the information displays of Last.fm and the fundamental fact that data from the database can be sorted and retrieved in different ways, allowing for seemingly limitless permutations. After that, the three sets of rankings (November 2014, June 2015 and December 2015) were compared to better understand the dynamics of the rankings of similar artists. Note that the length of each similarity ranking retrievable for each individual artist is 250 artists. To quantify the dynamics of the rankings, it is necessary to ascertain how many

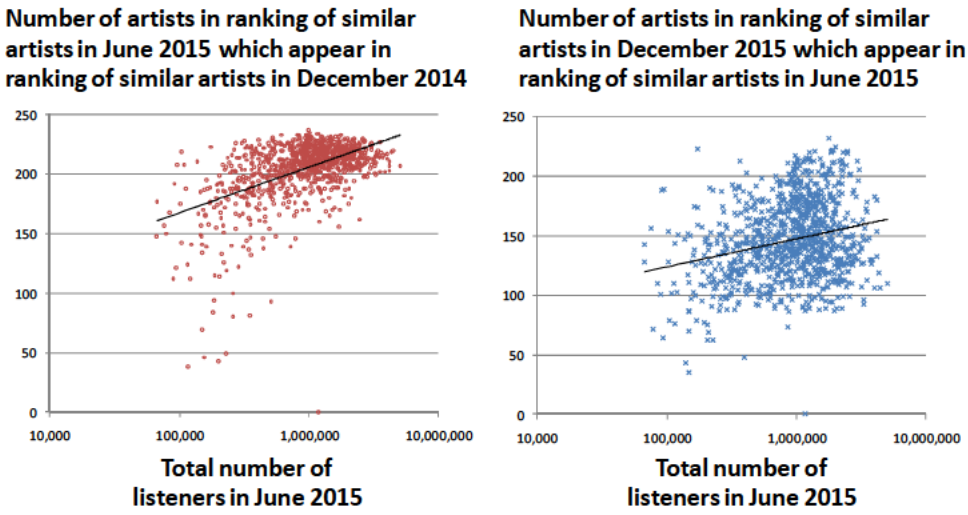
artists on each ranking of similar artists in June 2015/December 2015 appeared on corresponding rankings of similar artists in November 2014/June 2015. These variables are used to characterize the dynamic nature or the stability of the rankings over the two periods between November 2014 and June 2015, and between June 2015 and December 2015.

Figure 2 illustrates histograms for the two variables. The dynamics between the two periods are strikingly different: changes to the rankings of similar artists over the earlier period are much more stable than over the latter period. This is because the earlier period entails the business-as-usual scenario of Last.fm without significant changes to its operation. Arguably, this is a lot more stable than one might expect. The seamless updatability of the database still occurs during the business-as-usual scenario and this suggests that rankings of similar artists can be highly fluid and unstable as they are updated with a new stream of listening and tag data (Kallinikos et al., 2013). However, this is not what is found. In Figure 2, the number of artists in a ranking of similar artists in June 2015, who also appear in a ranking of similar artists in November 2014, stands at more than 200 for the majority of popular artists. These, arguably, recount the self-fulfilling prophecy mechanism as discussed by Espeland and Sauder (2007), as users react to similarity rankings and generate listening data in such a way that artist similarity rankings stabilize during a business-as-usual scenario. This is because artist similarity rankings are used as input in the assembly of personalized recommendations. Therefore, the assumption of similarity in accordance with computed artist similarity ranking is reinforced as users accept these recommendations. On the other hand, the rankings of similar artists over the latter period, between June 2015 and December 2015, is much more unstable. This, arguably, reflects changes to the recommendation algorithms (and also presumably the algorithms underlying artist similarity rankings) with significant changes to Last.fm in September 2015. After all, algorithms which assemble digital information

displays are also digital artifacts; therefore, they can always be changed as they are characterized by openness (Kallinikos et al., 2013).

Figure 8 plots the two variables (the number of artists in a ranking of similar artists in June 2015/December 2015 which also appear in a ranking of similar artists in November 2014/June 2015) on the y-axis and the number of total listeners on the x-axis (logarithmic scale). A positive correlation can be clearly detected. The values of the variables will tend to be higher for artists with a larger number of listeners. For artists with around one hundred thousand total listeners, the figures can be as low as 50. However, for artists with a number of total listeners around one million, the figures can be as high as 200. All this suggests that rankings of similar artists tend to stabilize as more listeners listen to corresponding artists or as time goes by, assuming ceteris paribus. However, note that the trend line for the earlier period is much higher than that of the latter period, reflecting a more stable ranking of similar artists in general. Again, this trend should not be surprising if one recognizes that the collaborative filtering recommender algorithm which Last.fm utilizes suffers from a popularity bias in which more popular items are recommended more frequently. Celma and Cano (2008) demonstrate this in the case of Last.fm itself. Therefore, the force derived from the self-fulfilling prophecy mechanism as discussed by Espeland and Sauder (2007) will be stronger for more popular items.

Figure 8: Dynamic of similarity ranking and total number of listeners



So, one way that artist similarity ranking casts influences upon user choices and behavior is indirectly through personalized recommendations. Another way is more directly through the browsing activities of users. As already discussed, Last.fm also uses artist similarity rankings to organize music on its websites and users are free to browse through these rankings to discover music. To investigate this, the number of incoming links from artist similarity rankings to different artists and music demand ought to be examined. Oestreicher-Singer and Sundararajan (2012b) empirically verify in a static case that the demand will be higher for products of Amazon with more incoming hyperlinks from its co-purchase network. Others replicate the empirical analysis of Oestreicher-Singer and Sundararajan (2012b) and verify that music demand will be higher for artists with more incoming hyperlinks from artist similarity rankings in the case of Last.fm. So, it would be reasonable to postulate that in this case, entailing comparisons in time, demand (as measured by number of listeners) for music from artists who experience an increase in the number of incoming hyperlinks from artist similarity rankings should also increase.

An important implication of the dynamics of artist similarity rankings for user choices and behavior thus emerges. During the business-as-usual scenario (between November 2014 and

June 2015) in which the database of Last.fm was merely updated with a new stream of social data, rankings of similar artists were very stable (Figure 7). Moreover, they tended to stabilize as time passed or as the number of listeners increased (Figure 8). On the other hand, changes to the rankings of similar artists impact on user choices and behavior as they have implications on the demand for the music of different artists (Oestreicher-Singer and Sundararajan, 2012b). All of this suggests that the impact of the kind of information displays investigated here (i.e. artist similarity rankings) is to stabilize music demand for different artists as generated by browsing activities over time, as the rankings of similar artists stabilize during the business-as-usual scenario.

However, it is important to note that this business-as-usual scenario and stabilizing of dynamics are not permanent. Once in a while, there can be significant changes to the operation of Last.fm (presumably also with changes to the underlying recommendation algorithms) as for example, when the business goal/plan of Last.fm changed. This happened in September 2015 and its impact was reflected in changes between June 2015 and December 2015. Such an event acts to disturb the stabilizing dynamics of the rankings of similar artists as reflected in Figure 2, whereby rankings of similar artists become relatively unstable. This will also destabilize the demand for different artists as generated through browsing activities as changes to incoming links from rankings of similar artists are associated with changes in music demand (Oestreicher-Singer and Sundararajan, 2012b). Thereafter, it is to be expected that Last.fm would enter a business-as-usual scenario once again, with stabilizing rankings of similar artists and of the music demand for different artists as the mechanics discussed by Espeland and Sauder (2007) came into effect. However, again, stabilizing dynamics do not last as a recommendation algorithm is, after all, a digital artifact characterized by openness and can always be altered.

Conclusions and discussion

As digital information assembled by social media platforms is unpacked, three theoretical claims about digital information displays as assembled by social media platforms emerge. First, ranking is supposedly the dominant strategy used by social media platforms to display information. This is because social media platforms use social data to assemble information displays and behavioral (quantitative, countable) data is of most importance to social media platforms (Alaimo and Kallinikos, 2016). Therefore, ranking can be readily assembled as items can be ordered by quantitative data, raw behavioral data or computed 'scores'. Second, navigation through digital information displays assembled by social media platforms ought to be a highly interactive experience. This is because the digital information displays of social media platforms are supported by a database (of social data) at the backend and an algorithm that can slice and dice data in the database in different ways, allowing for a seemingly limitless permutation of items to be displayed. Hence, alternative pathways of exploration can always be enacted for different users to pursue their different and evolving interests (Manovich, 2001). Third, each pathway of exploration in accordance with the digital information displays of social media platforms ought to be highly fluid and unstable. This is because, after all, this information display is a digital artifact, which is supposedly in constant flux (Kallinikos et al., 2013). Further, it is assembled by a database and an algorithm which are also digital artifacts that change over time. For example, a database is being constantly updated with a new stream of social data and an algorithm is characterized by openness and can always be changed. This has culminated in interaction systems as discussed by Wegner (1997) for content discovery.

A key contribution of this paper is to discuss how social data is combined with algorithms to create information displays of social media platforms and arrive at the three theoretical claims, while assessing the information display strategy of a real world social media

platform, i.e. Last.fm, against these theoretical claims. Results determined that ranking is indeed the dominant strategy to display information utilized by social media platforms. Last.fm utilizes countless charts to display music. Further, the display of artist similarity and personalized music recommendations can also be considered as ranking. Using database technology, Last.fm manages to assemble this diverse set of information displays for its users to navigate through music. Thus, navigation through the digital information displays as assembled by Last.fm is a highly interactive experience.

The assessment that leads to some surprise is related to the fluidity and instability of digital information displays. Here, artist similarity rankings are assessed over time. The type of information display as assembled by Last.fm is assessed because this is of central importance to the operation of Last.fm as it acts as an input into the assembly of personalized music recommendations. Surprisingly, artist similarity rankings appear to be more stable than initially expected. This demonstrates that the theory of digital objects (Kallinikos et al., 2013) alone cannot fully explain the dynamics of digital information displays. Instead, the theory of digital objects must be combined with the theory of information displays (Kleinmuntz and Schkade, 1993), particularly when related to ranking (Espeland and Sauder, 2007) to fully explain the dynamics. Similarity artist rankings appear to be stable during the business-as-usual scenario of Last.fm, whereby its database is simply updated with a new stream of social data because its dynamics exhibit a self-fulfilling prophecy as documented by Espeland and Sauder (2007). However, this stabilizing dynamic does not last as the algorithm which assembles the information display of Last.fm is also a digital artifact which is characterized by openness and can always be altered. However, the stabilizing dynamics should supposedly kick in again as Last.fm enters a business-as-usual scenario.

This leads to potentially interesting implications for music demand which will be higher for artists with more incoming hyperlinks from artist similarity rankings (Oestreicher-Singer and

Sundararajan, 2012b). Stabilizing artist similarity rankings during the business-as-usual scenario of Last.fm implies that music demand ought also to be stabilizing. However, this will not last as digital artifacts are in constant flux. After all, the algorithm which assembles the information display of Last.fm is characterized by openness and can always be altered as the business plan/goal of Last.fm changes. This will inevitably bring instability to artist similarity rankings, and so also for music demand of different artists on Last.fm.

(How) Does Data-based Music Discovery Work?

Akarapat Charoenpanich, LSE

Aleksi Aaltonen, Warwick Business School

Abstract

This paper analyses a new type of business operations that mediate the production and consumption of music. The online environment has largely abolished constraints on the variety of music that can be economically distributed, but, at the same time, it reveals another problem. How do people learn what music items they want to listen to? In the music industry, the product space consists of thousands of artists, songs and albums, and is expanding rapidly. More effective forms of music discovery could therefore create considerable new value by allowing people to listen to music that better matches their taste. We analyse data from the Last.fm music discovery service that deploys a collaborative filtering recommender system and social media features to aid music discovery. The analysis finds evidence that the new form of music discovery is valuable to consumers, yet it is relatively less important than an opportunity to listen to music for free. The findings lead us to discuss how the nature of analytical problem and product space, consumer taste, and social media features shape the potential value of created by big data.

Introduction

In this paper, we study data-based business operations that mediate the production and consumption of music in the digital ecosystem. We define music discovery as a process by which people identify new music items that are subsequently incorporated into individual music consumption. Music discovery can happen in many different ways. For instance, consumers may actively browse racks of CDs, search online catalogues, or be guided by social cues and recommendations from their environment. Importantly, people often do not know what they want to listen to until they have actually started listening to it, which gives music discovery often an exploratory nature. It differs considerably from known-item type seeking, that is, locating items that they already know (Morville and Rosenfeld, 2006). The limitations of physical distribution channels have traditionally pushed music consumption toward the most popular artists, and there is a hope that new digital platforms could unleash the potential of niche items in the long tail of consumer demand (Anderson 2006; Celma 2008).

Assuming that people's 'true' music taste is more diverse than what traditionally narrow distribution channels have been able to serve, more effective forms of music discovery can create considerable new value by allowing people to listen to music that better matches their taste. Our empirical analysis focuses on an online service that musters social consumption data to provide personalized music recommendations. We analyse a dataset retrieved from Last.fm music discovery service that deploys a collaborative filtering recommender system and social media features to aid music discovery. The company was founded in 2002 and is one of the most popular services of its kind today. Last.fm users submit listening data from over 600 playback applications, services and devices to receive recommendations for further music items. Whether these recommendations are valuable to consumers is, however, an empirical question that goes to the heart of a business built on data and data analytics

(Aaltonen and Tempini, 2014). We ask the following research question: *Does the new form of data-based music discovery provide value to consumers?*

To answer such a question, one needs to be able to separate music discovery and its value from the value of sheer music acquisition. Consumers can undoubtedly find value in dirt-cheap music streaming, but do they find the new form of music discovery *per se* valuable? The answer has important managerial implications and can help us better understand data-based innovations and business models. We develop a theoretical model of music discovery and consumption, and harness changes in Last.fm consumer offering to separate music acquisition from music discovery.

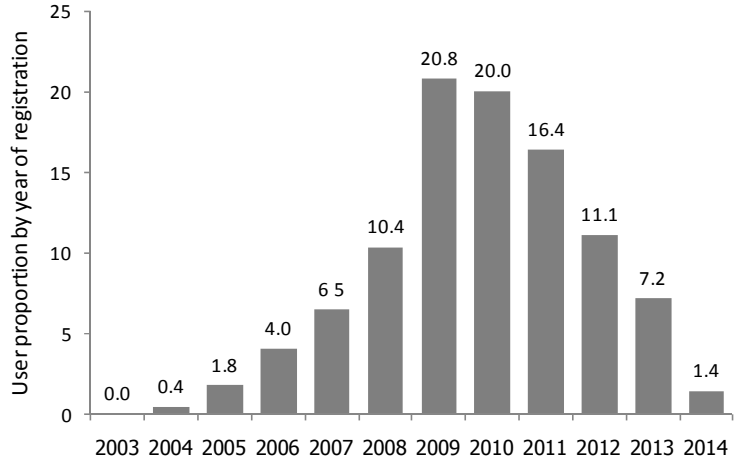
The analysis uses the amount of music consumption as an indicator that a consumer finds the service worth using (and hence valuable) in a competitive market environment (Oestreicher-Singer and Zalmanson, 2013). The main dependent variable is thus not a direct measure of commercial success but an important prerequisite for generating revenues and increasing the valuation of online businesses (Brynjolfsson et al., 2013). We find evidence that the new form of music discovery is valuable to consumers, yet it is relatively less important than an opportunity to listen to music for free. Also, whether the value of music recommendations can be captured to support a viable business is a different matter. We will return to these matters in the discussion of findings, which call for more attention to the underlying mechanisms of value creation and capture for data-based business.

Music discovery through Last.fm

Last.fm is one of the oldest and most popular online music discovery services. The service collects music listening and social data from over 600 playback applications, services and devices to create personalized music recommendations. Since the inception of Last.fm in 2002 to early 2009, users could stream free music directly from the service. This undoubtedly

contributed to the rapid growth seen in Figure 1 below. In April 2009, the company limited free streaming to the US, UK and Germany, citing its inability to recover music licensing fees from advertising. Users in other countries were then required to pay a subscription fee for streaming. Over the last five years, Last.fm has gradually wound down all streaming operations, focusing its business exclusively on music discovery. The service is based on continuously amassing user-generated music consumption data from the digital ecosystem, and carefully distilling them into personalized music recommendations that are supported by various social media features. The users are encouraged to stream music provided by partners such as Spotify and YouTube.

Figure 1: Last.fm User Growth (user proportion by the year of registration)



In the following brief literature review, we describe three factors that shape data-based music discovery and consumption through Last.fm. At the heart of the Last.fm there is a collaborative filtering recommender system (CFRS) that is a popular approach to providing product recommendations. Second, raw data that feeds the Last.fm CFRS is retrieved from hundreds of different sources and must be cleansed and amended with appropriate metadata so that the data becomes a reliable resource for computational processing. Finally, music discovery and consumption are highly social activities. The service allows users to interact

around music items, artists and events, which can both enhance music discovery and make users more committed to the service.

Collaborative filtering recommender system

People reveal their music preferences by listening more often and repeatedly to music they like. Playback counts can therefore be used as ratings data for a collaborative filtering recommender system (Ekstrand et al. 2010). The CFRS uses the data to construct a collective similarity network between music items, maps an individual user's preferences to the network, and, finally, produces recommendations regarding nearby products in the network. Anderson (2006) believes the approach could increase demand for niche products and there are some highly successful applications such as Amazon's "Customers Who Bought This Item Also Bought" feature that drive sales by automatically created recommendations. However, CFRS alone is not a panacea for finding relevant products. This may be due to the lack of suitable, high quality data but also relates to the specific nature of items to be recommended. Goldenberg et al. (2012) point out that the fact that the underlying product network is constructed simply on similarities between items can be particularly problematic for music.

Music recommendations should be both novel and relevant (Celma and Lamere 2011). Recommending *Nine Inch Nails* to a *Prince* fan may be novel but probably not very relevant whereas *Michael Jackson* would be perhaps relevant but most likely redundant. There are different opinions on what kind of dynamics recommender systems stimulate in general and specifically in the context of music business. Oestreicher-Singer and Sundararajan's (2012b) study of Amazon.com shows that the structure of a product network does not only reflect peoples' past purchases but it can also affect subsequent demand for products. Computational recommendations may not thus be a straightforward reflection of similarities between items in the product space but a part of complex, dynamic process that, at least to a degree, shapes

what it is supposed to reflect. On the one hand, Celma and Cano (2008) find that the Last.fm similarity network suffers from a popularity bias and go on to argue that this may be an inherent problem associated with the use of social data to organize content (see also Hosanagar et al., 2013). In Last.fm, the play count of artists is strongly correlated with the play count of other similar artists. Popular artists are more likely to act as hubs within the similarity network, while less popular artist are less likely to be recommended even if discovering those long tail items could often be most valuable. On the other hand, Levy and Bosteels (2010) who are Last.fm employees defend the service against popularity bias criticism using an internal dataset.

Metadata infrastructure

The CFRS can only work if it can reliably identify two pieces of music content as the same or different items, and associate them with a correct description. It is important to bear in mind that the system does not operate directly on music content but analyses its metadata that is an important infrastructure for many operations in the industry (Jannach et al. 2011; Brookes, 2014a; 2014b). Metadata is a description of a resource. It informs about the structure and content of a bundle of data that may represent a song, photograph, database or any other digital artefact (Kallinikos et al., 2013). Critical metadata used to be permanently printed on top of vinyls and CDs, whereas digital music files are much more loosely coupled with their metadata and may easily lose it. Without appropriate metadata, it is impossible to manage music content and its copyrights on a commercial scale, or even find anything from tens of millions songs available through online music services. At the moment, there is no authoritative, institutionally controlled source of music metadata and a lot of music circulates in the digital ecosystem with partial, inconsistent and simply incorrect metadata.

Humans can deal with small inconsistencies and errors in music metadata, but these factors pose considerable problems for computational processes that underpin CFRS and music

business in general. Last.fm needs to be able to analyse music consumptions across a broad range of services and devices that source metadata from at least five different vendors all imposing their own metadata standards. Furthermore, peer-to-peer file sharing services allow user-generated ID3 metadata tags to propagate throughout the digital ecosystem as the tags are not controlled by anybody (Morris, 2012; Brookes, 2014a; 2014b). Even more importantly, no metadata is useful unless it can be associated with the right content. A music identifier is a unique token that ties metadata to a music item, and allows identifying two pieces of music content as the same or different items. The lack of reliable identifiers makes it difficult to calculate play counts, compute recommendations and, in general, to ensure items are presented with the correct description of the music content. Last.fm relies mainly on the combination of artist name and song name as the identifier, but given the poor quality of existing metadata the approach is far from perfect. For instance, the company has found that there are over 100 ways to spell the artist name – song name combination *Guns N' Roses – Knockin' on Heaven's Door*.

Social media features

Music discovery and consumption are typically highly social activities. Last.fm users often begin as mere consumers of recommendations but may eventually start to participate more intensively, for example, by creating and organizing content, participating in discussions, and even become informal leaders in the community (Oestreicher-Singer and Zalmanson, 2013). This is important because socially engaged users have been found to be more likely to pay for Last.fm as well as other services (Fullerton, 2003; 2005; Oestreicher-Singer and Zalmanson, 2013), and once a user subscribes to a premium service, the likelihood that his or her friends subscribes increases (Bapna and Umyarov, 2012).

Social media engagement can mitigate the shortcomings of CFRS in providing valuable music recommendations. Recommendations that are based on a similarity network

constructed from user data can be too successful in connecting similar products together and, arguably, biased toward popular items. This can easily render the output from CFRS less useful for the users. The integration of a similarity (product) network with a social network into a dual network approach can alleviate the problem. Users create idiosyncratic links to the similarity network as they participate in social media activities by posting comments, writing reviews, tagging content, etc. These links can be seen as their personal recommendations and ways of grouping products, which can complement similarity network and help users to discover more relevant items (Chen et al., 2010; Goldenberg et al., 2012).

Data collection and the dataset

We collect social consumption data from Last.fm to evaluate a theoretical model of music discovery and consumption. The data are mainly retrieved via the Last.fm Application Programming Interface (API) without personally identifying information. The only exception to this is the username that may sometimes represent the real identity of a sample user. Usernames are not included anywhere in the reported findings. The construction of a dataset for statistical analysis involves three steps: 1) identifying a representative sample of Last.fm users, 2) retrieving data for each user in the sample, and 3) assembling the dataset with variables that operationalize music discovery and consumption.

We apply a rejection sampling method proposed by Gjoka et al. (2010) to retrieve a representative sample of users. Each Last.fm user has a unique positive integer as his or her identifier. The identifiers are generally assigned so that a user who registers later will receive a larger number, and the entire user population should comfortably fall within a range of 1 and 100,000,000. We draw a random integer from the range and query Last.fm for data by using the number as the user identifier. We repeat the procedure storing raw data until we end

up with a random sample of 12,839 users, which is deemed large enough for regression analysis

Dataset construction

We retrieve five types of data for each user in the sample and assemble them to a panel dataset that traces users through time along several variables. We divide the temporal dimension of the panel dataset into yearly increments, which allows us to separate the impact of changes to Last.fm consumer offering without breaking the dataset into too small subsamples. The dataset describes individual users with five main variables that allow us to unpack the impact of data analytics and social media features on music discovery and on music consumption.

Playcount measures the amount of music consumption. The variable is based on listening event data that represent the playback or streaming of individual songs. The data include the title, artist name and time for each song a user has listened to. We simply count the number of annual listening events per user, which is the sole input to the variable.

Listening concentration measures the relative success of music discovery. Chen et al. (2010) found that after a successful discovery there is often a burst of listening as the user keeps listening to music from the same artist for a period of time in Last.fm. Consequently, we assume that a more concentrated listening profile at the artist level signals more successful music discovery. We use the listening event data to compute a Herfindahl Index (HI) as an operational measure of concentration (Benkler, 2006; Kwoka 1985; Rhoades, 1993). Note that we also normalize our HI to ensure we can better use it to compare listening concentration of different users across time. The HI is described in more detail in Statistical Appendix.

Friends measure social media engagement in Last.fm. We retrieve a friend list for each user, which represents social relationships that the user has actively acknowledged at the end of the observation period. We also retrieve the time of each public communication between the user and his or her friends. Using these two types of data, we construct a proxy variable that traces the number of friends at different points in time by assuming that the time at which each connection of friendship is established coincides with the time at which the users communicated for the first time in Last.fm. This should give a reasonably accurate, lower bound estimate of the number of friends at different times, since our panel dataset observes the temporal dimension only at the annual level. Although users can communicate without adding each other to the list of friends, by combining the two types of data we intend to increase the reliability of the measure and ensure that we capture positive emotional relationships within the user community (Chmiel et al., 2011).

Auto-corrections measure the quality of metadata that makes personalized music recommendations possible. Last.fm introduced in January 2009 a system that can automatically correct artist and song names, and therefore counter problems stemming from incorrect music identifiers circulating in the digital ecosystem. We retrieve all auto-correction mappings applied to the listening events of our sample users. The mappings consist of artist names submitted by users that are deemed incorrect, and the correct names to which they are mapped to. We construct a proxy variable to trace the number of corrections made to the listening data of each user over time. This is done by estimating the number of artist names that have been corrected for each user by comparing auto-correction mappings with the listening events of each user. We assume for certain names on auto-correction mapping to appear for the first time when those names appear on listening data of users for the first time. Since we can only retrieve listening events whose metadata has been already corrected, we do not know the original metadata that the user submitted to Last.fm.

Past listening similarity measures the utilization of data analytics. Ideally, we would like to observe actual personal recommendations produced by the CFRS through time. However, since no such data is easily available, our next best option is to analyze the similarity of current listening to past listening. This is because the logic of CFRS is to use a product similarity network to recommend products similar to those that the user has ranked highly in the past. Therefore, we expect the listening events to be relatively more similar to the music which the user has listened to previously if the user relies on the recommender system to discover new music. Although actual personal recommendations produced by CFRS cannot be observed, we can still study the underlying product similarity network. The approach has been previously implemented by, for example, Celma and Cano (2008).

We compute the proxy variable for past listening similarity. Since a similarity network between products is known to be usually relatively stable over time (Konstan and Riedl, 2012), we simply use a static network at the time of data retrieval retrospectively for all calculations. First, we pull a list of top 50 similar artists for each listening event in the sample. We then identify different artists that user has listened to previously in relation to each listening event. Finally, we sum the number of those different artists associated with each listening event, and average these values annually. This is done by computing the value at each listening event for a year and taking their arithmetic mean. Note that we calculate past listening similarity for users with more than 10,000 listening events by randomly selecting only 10,000 events for calculation. This improves computational speed, while the accuracy of the calculation is also ensured since it is based upon random selection.

Table 1: Dataset Construction (time series data marked with *)

Variable	Data	Concept
PLAYCOUNT	Listening events*	Playcount is the main dependent variable that measures the amount of music consumption.
LISTENING CONCENTRATION	Listening events*	Artist-level concentration of music consumption measures the success of music discovery.
FRIENDS	Friend list Time of communication between users*	The number of friends is a proxy for the use of social media features.
AUTO-CORRECTIONS	Auto-correction mappings Listening events*	The number of auto-corrections is a proxy for the quality of underlying metadata
PAST LISTENING SIMILARITY	Lists of 50 most similar artists Listening events*	Past listening similarity is a proxy for the use of data analytics.

Table 1 summarizes the data and concepts that are used to construct the five main variables for path analysis. Out of the five types of retrieved data, listening events and the time of communication between users are time series while auto-correction mappings, friend lists and the lists of 50 most similar artist represent the situation at the end of the observation period 30 June 2014. FRIENDS, AUTO-CORRECTIONS and PAST LISTENING SIMILARITY variables are therefore constructed as retrospective estimates in our panel dataset by computing proxies for them. Finally, users can opt to hide their listening event data and to turn the auto-correction feature off, but this is rare in practice.

Descriptive statistics

In our sample of 12,839 Last.fm users, 57 per cent have submitted at least one listening event, 22 per cent have had their music metadata corrected by the auto-correction system, 9 per cent have at least one person in their friend list, and 5 per cent have communicated publicly with their friends through Last.fm. Listening data for the sample users amounts to 18,804,414 events that, by construction, is the sample users' total aggregate playcount. These listening

events include 221,614 different artist names. For each of the artists, we retrieve the list of 50 most similar artists. Since some of the variables cannot be calculated for a user that has zero playcount, we have dropped such users during such time periods from the table and any further analyses. We also drop users who listen to only one artist, since normalized HI cannot be computed for them.

Table 2: Descriptive Statistics for the Five Variables

Variable	Mean	Median	Std. deviation
PLAYCOUNT	1659.6	101.0	6199.2
LISTENING CONCENTRATION	317.9	93.4	752.5
FRIENDS	1.2	0.0	6.8
AUTO-CORRECTIONS	13.6	1.0	40.9
PAST LISTENING SIMILARITY	7.3	5.1	7.2

Table 2 reports descriptive statistics for the five main variables in our dataset. The means for each variable are considerably higher than medians that suggest highly skewed distributions, which is common in social data (Shirky 2008). Hence, we transform the variables logarithmically and then compute a correlation matrix presented in Statistical Appendix. We find that the variables are significantly correlated and include a variance inflation factor (VIF) check for potential multicollinearity problems. Note that although mean and median of friends is relatively low (since only 5% of our users have publicly communicated with their friends), almost 20% of our data points register positive number of friends.

Table 3: Annual Means for the Five Variables in the Panel Dataset

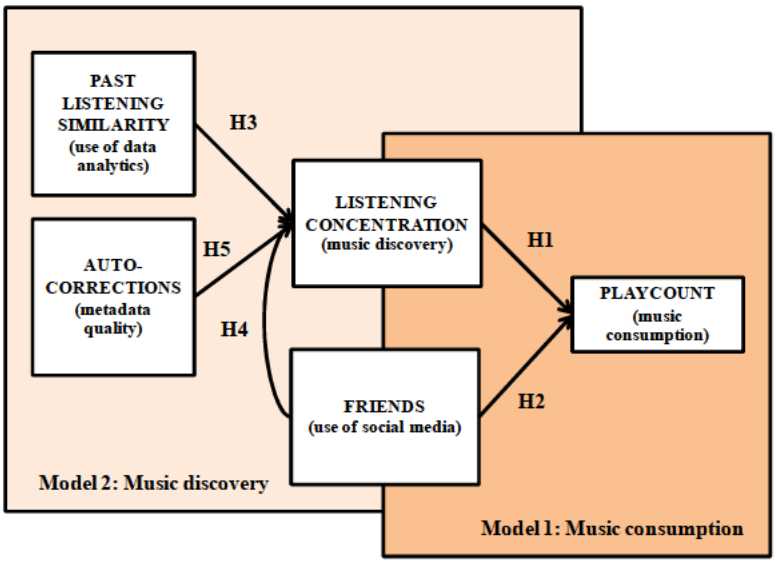
Year	PLAYCOUNT	LISTENING CONCENTRATION	AUTO-CORRECTIONS	PAST LISTENING SIMILARITY	FRIENDS
2005	3508.9	434.2	6.9	4.2	0.0
2006	2308.5	415.7	8.4	4.7	0.4
2007	2166.2	458.3	9.4	5.6	0.8
2008	1932.8	431.9	11.0	6.5	1.2
2009	1310.0	250.5	9.1	5.8	0.8
2010	1248.1	243.8	9.7	6.4	0.9
2011	1281.7	247.2	10.5	7.4	1.0
2012	2278.6	312.5	18.4	8.7	1.8
2013	2332.9	427.5	25.6	10.2	2.0
2014	1747.7	549.4	41.8	13.3	3.3

Figure 1 shows that the number of sample users peaks in 2009 and declines thereafter, which matches the overall pattern of new users in Last.fm. The peak coincides with the changes to the Last.fm consumer offering indicating that these changes may have had a significant impact on Last.fm usage. Table 3 presents annual means for the five key variables. It is worth pointing out that the auto-correction variable gets positive values even before the system was activated in 2009. This is because we rely on a proxy variable and do not know the time when a particular correction was first applied to user-submitted metadata. We compensate for this problem in our main estimations by running our estimation twice, before and after 2009. Differences between the two estimations allow us to make inferences about the impact of auto-correction system.

Theoretical model

The empirical analysis consists of estimating two equations that capture together eight hypotheses on how Last.fm works. Five of these are captured in Figure 2 that shows a path diagram for a theoretical model of music discovery and consumption. The first equation (Model 1, note that we read the diagram from right to left) estimates factors that influence music consumption, while the second equation (Model 2) opens up music discovery. As the model itself is relatively straightforward mapping of causal relationships found in the previous literature, our main interest is the changes before and after the major changes to the consumer offering in 2009.

Figure 2: Path Diagram of Music Discovery and Consumption



We expect *ceteris paribus* that users consume more music (i.e. more number of tracks users listen to) if they are able to discover interesting artists and engage social media around music. This is captured in Model 1. Bateman et al. (2011) show that online participation is directly linked to commitment, as defined by organizational commitment theory. There are three types of commitment: continuance, affective and normative commitment.

H1: Better music discovery leads to more music consumption (continuance commitment)

H2: More intensive social media engagement leads to more music consumption (affective/normative commitment)

Model 2 opens up data-based operations underpinning music discovery in more detail. We expect users to be more successful in discovering music (i.e. discover artists users like and start to listen to those artists intensively) if they use the recommendations (Chen et al., 2010), engage in social media activities (Goldberger et al., 2012), and their music items have correct metadata (Brookes, 2014a; 2014b). By definition, more music discovery will lead to more concentration of listening data in terms of artists. We assume the following hypothetical relationships in Model 2.

H3: More utilization of data analytics leads to better music discovery (dual network)

H4: More intensive social media engagement leads to better music discovery (dual network)

H5: Better metadata quality leads to better music discovery (metadata problem and information infrastructure)

Closing down music streaming from 2009 onward has had a major impact on Last.fm users, which may cast a direct negative impact not only on both music consumption but also, interestingly, on music discovery. Slowing growth and decline in users number as a result of closing down streaming operations 2009 onward may affect music discovery since Last.fm gets less timely data on new artists and songs. Therefore, we estimate the models separately

for periods before and after 2009 to formulate two additional hypotheses that isolate the effect of business model change.

H6: The intercept term for Model 1 (music consumption) is lower after 2009 as compared to estimation before 2009 (changes of consumer offerings and continuance commitments)

H7: The intercept term for Model 2 (music discovery) is lower after 2009 as compared to estimation before 2009 (changes of consumer offerings and slower growth of social data)

Since the auto-correction variable gets positive values even before the system was activated in 2009, running the estimation separately before and after 2009 also allows us to better assess the effect of the auto-correction system. One of the main effects of the auto-correction is to lump data that was previously associated with different artists together, increasing, naturally, listening concentration. Because we only observe listening events that have already been mapped by the auto-correction system, effect of auto-correction variable should be positive and significant for music discovery even before 2009. This purely illustrates the lumping effect. We expect the positive effect associated with the estimation after 2009 to be even stronger. And, here, the incremental positive effect can be interpreted as the positive impact of metadata quality upon music discovery.

H8: Positive association between auto-correction variable and listening concentration is stronger for estimation after 2009 as compared to estimation before 2009

Findings from a path analysis

We conduct a path analysis by estimating the two models (music consumption and discovery) for two time periods (before 2009 vs. 2009 and onward) using simple ordinary least square estimation. For this purpose, we need to make a few additional adjustments to our panel dataset. First, we pool our panel dataset across time since, here, we run our estimation based on pooled panel data. Second, we include only data points with at least one listening event, since some of the variables can only be computed for users with a positive PLAYCOUNT value. Also, we drop data points, whereby users listen to only one artist since we cannot compute normalized HI for those data points. Third, we apply a logarithmic transformation to adjust the highly skewed distributions of the five main variables. After that, we apply ordinary least square estimation to estimate the following two equations for the two time periods.

(1) $\text{Log (PLAYCOUNT or music consumption)} =$

$\alpha_0 + \alpha_1 \text{Log (LISTENING CONCENTRATION or music discovery)} + \alpha_2 \text{Log (FRIENDS or use of social media)}$

(2) $\text{Log(LISTENING CONCENTRATION or music discovery)} =$

$\alpha_4 + \alpha_5 (\text{PAST LISTENING SIMILARITY or use of data analytics}) + \alpha_6 \text{Log (FRIENDS or use of social media)} +$

$\alpha_7 \text{Log(AUTOCORRECTION or metadata quality)}$

Table 4 and Table 5 report the estimation result for the two models during the two periods. All coefficients show hypothesized signs and most of them are statistically significant at one per cent level ($\text{Sig} \leq 0.01$). The tables compare the relative importance of different factors for music discovery and music consumption in Last.fm for the two time periods. Further, policy changes from 2009 onward have a significant negative effect on music consumption and, more interestingly, on music discovery, as reflected in changes to the intercept terms for estimations before and after 2009.

Music discovery (Model 2)

Table 4: The Impact of Data Analytics, Data Quality and Social Media Engagement on Music Discovery

Coefficients[a]

Dummy_2009onward		Unstandardized Coefficients		Standardized Coefficients	T	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
0	(Constant)	3.656	.080		45.938	0.000		
	log_analytics	.397	.051	.175	7.854	.000	.504	1.986
	log_dataquality	.095	.029	.077	3.330	.001	.467	2.142
	log_friends	.076	.041	.034	1.865	.062	.729	1.372
1	(Constant)	2.245	.050		45.329	0.000		
	log_analytics	.805	.030	.312	26.538	.000	.774	1.292
	log_dataquality	.328	.023	.178	14.247	.000	.685	1.459
	log_friends	.133	.054	.028	2.475	.013	.825	1.212

a. Dependent Variable: Log_HI_normalized

Table 4 reports the impact of data analytics, data quality and social media engagement on music discovery. The first four rows (Dummy_2009onward = 0) present the estimation for 2002–2008, while the last four rows present the estimation for 2009–2014 (Dummy_2009onward = 1). The variance inflation factor (VIF) values in the two last columns show that the findings are not subject to multicollinearity problems. We also test if the observed differences between the time periods are statistically significant.

We find that the use of music recommendations and data quality have considerable effects on music discovery, while social media engagement has only a weak effect. Most importantly, successful music discovery is expected to increase by 0.397 per cent (before 2009) and 0.805 per cent (2009 and after) against 1 per cent increase in the use of music recommendations. This shows that consumers find data-based recommendations useful. Also, we find that the difference between the time periods is statistically significant, which means that the company is able to make its recommender system progressively more effective. Furthermore, music discovery is expected to increase by 0.095 per cent (before 2009) and 0.328 (2009 and after) against 1 per cent increase in the data quality. The difference between the time periods is statistically significant and results most probably from the activation of auto-correction system in 2009. At the same time, social media engagement has a very weak effect on music discovery as the latter is expected to increase only by 0.076 per cent against 1 per cent increase in social media engagement (the difference between time periods was found statistically insignificant). Changes to consumer offerings also cast significant influences upon music discovery as the intercept term of the estimation for the 2009 and after time period is much lower than that for the before 2009 time period. The difference is statistically significant and indicates that music discovery is expected to decrease by as much as 75.6% because of the policy changes. The dependent variable of Model 2, music discovery, enters into Model 1 as an explanatory variable. The variables in the equation may therefore cast an indirect effect on music consumption.

Music consumption (Model 1)

Table 5: The Impact of Music Discovery and Social Media Engagement on Music Consumption

Coefficients[a]

Dummy_2009onward		Unstandardized Coefficients		Standardized Coefficients	T	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
0	(Constant)	3.639	.087		41.992	0.000		
	Log_HI_normalized	.377	.017	.308	22.044	.000	.977	1.023
	Log_friends	1.052	.038	.388	27.730	.000	.977	1.023
1	(Constant)	2.375	.040		59.422	0.000		
	Log_HI_normalized	.465	.009	.477	51.338	0.000	.967	1.034
	Log_friends	1.360	.043	.297	31.957	.000	.967	1.034

a. Dependent Variable: log_playcount_annual

Table 5 reports the impact of music discovery and social media engagement on music consumption. The specification of the estimation is nearly identical to Model 2 above, but since social media engagement is an independent variable for our Model 2, it also has an additional indirect effect on music consumption through music discovery. The indirect effect is, however, relatively small. Again, the first three rows (Dummy_2009onward = 0) present the estimation for 2002–2008, while the last three rows present the estimation for 2009–2014 (Dummy_2009onward = 1). The variance inflation factor (VIF) show that the findings are not subject to multicollinearity problems, and we also test if the observed differences between the time periods are statistically significant.

In contrast to music discovery in Model 2, we find that social media engagement has a considerable impact on music consumption. Music consumption is expected to increase by 1.052 per cent (before 2009) and 1.360 per cent (2009 and after) against 1 per cent increase in social media engagement. Also, music discovery has a strong impact on music consumption that is expected to increase by 0.377 per cent (before 2009) and 0.465 per cent (2009 and after) against 1 per cent improvement in music discovery. The differences between time periods are statistically significant for both independent variables, which allows further interpretation. This indicates that Last.fm use has become increasingly focused on data-based music discovery that provides clear value to consumers, albeit with small indirect support from social media engagement. At the same time, social media features remain very important for user retention.

Policy changes regarding changes to consumer offerings also cast significant influences upon music consumption as intercept term of the estimation for the 2009 and after time period is much lower than that for the before 2009 time period. The difference is statistically significant and it translates to the direct negative impact of as much as 71.7%. Further, policy changes cast indirect negative impact upon music consumption through music discovery to reduce music consumption by further 35.2% (75.6% x 0.465). Henceforth the total effect of policy changes is to reduce music consumption by a whopping 81.3% $[1 - (1 - 71.7\%) \times (1 - 35.2\%)]$

Finally, Table 6 summarizes direct and indirect impact of data analytics, data quality, social media engagement and policy changes related to consumer offerings upon music consumption. It demonstrates that indirect effect of data analytics, metadata quality and social media via music discovery upon music consumption is relatively small. Although the direct

impact of use of social media upon music consumption is relatively large, such impact is still relatively small as compared to that related to policy changes. Henceforth, arguably, although new form of music discovery is valuable to consumers, the value is relatively modest compared to music acquisition, that is, music streaming.

Table 6. The Direct and Indirect Impact of Data Analytics, Data Quality, Social Media Engagement, and Policy Changes upon Music Consumption

	Before 2009			2009 and after		
	Direct	**Indirect**	**Total**	**Direct**	**Indirect**	**Total**
Increase in use of data analytics by 1%	n/a	+0.15%	+0.15%	n/a	+0.30%	+0.30%
Increase in metadata quality by 1%	n/a	+0.04%	+0.04%	n/a	+0.12%	+0.12%
Increase in use of social media by 1%	+1.05%	+0.03%	+1.08%	+1.36%	+0.05%	+1.41%
Changes to consumer offerings	n/a	n/a	n/a	-71.7%	-35.2%	-81.3%

Conclusion and discussion

We find evidence that the new form of music discovery and social media features are valuable to Last.fm users. However, value created by such operations need to be understood in context. The declining number of active users since 2009 suggests that the overall consumer value created by such operations is relatively modest compared to an opportunity to listen to music for free. Also, the value of data-based music discovery may not be perceived equally by all consumers but is likely more relevant to a specific type of music listeners. For instance, the Phoenix 2 UK project found that the proportion of music listeners who are enthusiasts is relatively small (Jennings 2007).

The findings raise questions whether big data supporting the venture can alone generate enough competitive advantage to sustain the business. In 2013, Last.fm made 2.1 million GBP loss, its revenues plummeted by 20 per cent, and the number of employees was halved (Sweney, 2014). Together with our finding that Last.fm depends heavily on social media features to retain its users beyond 2009, these observations call attention to key assumptions underpinning data-based music discovery business and, as we will elaborate below, big data innovations in general.

The new form of music discovery may well serve the needs of particular music enthusiasts whose music consumption is indeed limited by difficulties in findings interesting new music. Yet, for the majority of consumers this is probably not a major issue. Many people prefer to listen to *popular* music, that is, the very opposite of the long tail items. The concentration of music consumption on a relatively few popular items can look like a problem to some but it is also a testament to the social nature of music consumption. Popular music functions as a platform for socializing and makes it possible to share common experiences. Against this background, it is not surprising that significant amount of consumer value in Last.fm would seem to emerge from the use of its social features. This makes declining user numbers particularly problematic.

To an extent that the consumer value of Last.fm is created by social network externalities, loss of users numbers can perpetuate itself unless the service is able counter the loss of network externalities with increasingly successful music discovery. We find that music discovery has improved significantly over the years as Last.fm has enhanced its recommender engine and released new features such as the auto-correction system. At the same time, however, our findings show that the declining user base can also have a direct negative effect on digital music discovery. There are two reasons for this. First, it is less likely that the dual network is able to mitigate the problems of collaborative recommender

filtering system if there are less people on the platform. Second, the product space expands continuously with new music items that are new to all users. The less users there are submitting data, the longer it takes to capture enough ratings to incorporate new items. More generally, the importance of network externalities in social media is a well-known topic, and our study shows that the relative size of user base can also matter for the value of big data as a resource for product/service innovation.

Our analysis does not allow pinning down a causal model of data-based music discovery business, but it certainly opens up complexities involved in creating a big data business in a specific domain. These involve consumption patterns and the product space of a particular industry, the nature of analytical problem and its applicability to computational processing, and the role of social media and social data as a part of big data operations. In the case of music industry, the new form of data-based music discovery is valuable to some consumers and hence potentially a source of competitive advantage. At the same time, it may require sourcing data from a broader population to generate good recommendations. This leaves open a question, what is the benefit for those consumers?

Discussion about data-based businesses can become highly technical (e.g. Celma, 2008). Technical analyses are important and often insightful, but at the same time they may overlook other factors that are crucial for the successful operation of these businesses. Social media features (Oestriecher-Singer and Zalmanson, 2013; Goldenberg et al., 2012), industrial metadata (Brookes, 2014a; 2014b) and the nature of recommender systems are all important (Celma, 2008), but it is their interplay in a specific field of consumption that a company needs to understand if it is to reap sustained competitive advantage from products/services based on big data.

Statistical Appendix

Herfindahl Index

The Herfindahl Index (HI) is commonly used by competition economists to measure market concentration in mass media and in other types of markets (Benkler, 2006; Kwoka 1985; Rhoades, 1993). Here, we compute listening concentration of each user in accordance to HI and the formula is as followed.

$$HI = \sum_{i=1}^{n} \left(\frac{y_i}{\sum_{i=1}^{n} y_i} \right)^2 \times 10000, \text{ where}$$

y_i Total number of track of music associated with a particular artist within a particular year for a user

n Total number of different artist whom the user listen to within a particular year

$i \in \{1,2,3,\ldots,n\}$

Unfortunately, the problem with HI is that its lower bound depends on the number of different artist whom the user listen to. Since this varies considerably between the users in our sample, we use a normalized version of HI that ranges between 0 (extremely diverse) and 10,000 (extremely concentrated) regardless of the number of different artists the user has listened to. More precisely, it can be shown that HI would range between $1/n \times 10000$ and 10000 by its construct. Henceforth, sometimes the index is normalized with the following formula:

$$Normalized\ HI = \frac{HI - (1/n \times 10000)}{10000 - (1/n \times 10000)}$$

Correlation Matrix

	PLAY COUNT	LISTENING CONCENTRATION	AUTO-CORRECTIONS	PAST LISTENING SIMILARITY	FRIENDS
PLAYCOUNT	1	.500**	.728**	.721**	.449**
LISTENING CONCENTRATION	.500**	1	.331**	.377**	.198**
AUTO-CORRECTIONS	.728**	.331**	1	.576**	.514**
PAST LISTENING SIMILARITY	.721**	.377**	.576**	1	.362**
FRIENDS	.449**	.198**	.514**	.362**	1

*** Correlation is significant at the 0.01 level*

More than Networks: Social Media as Infrastructures

Cristina Alaimo, LSE

Akarapat Charoenpanich, LSE

Abstract

Despite the growing importance of social media technologies for the current development of the web economy (i.e. social buttons, APIs, etc.), the majority of IS contributions continue to see social media platforms predominantly as social networking sites. The static models of network analysis cannot capture the dynamics of the layered architecture of data exchange that underlies the complex infrastructuring of social media. They consequently risk missing what constitutes the novelty and specificity of these platforms: the distinct ways by which they produce, circulate and commercialize data and the new forms of interaction they propose.

We conduct an empirical study of Last.fm, a social media platform for music discovery, and we find that social media technologies are strongly associated with user listening activity, which results instead only tenuously linked to community participation. Our study lends support to the view of social media as infrastructures resting on integrated and layered social technologies that filter social participation, sustaining a continuous flow of social data across infrastructure layers and (increasingly) across business domains.

The primacy of social media technologies as generative mechanisms of social media networks suggests that firms cannot view social media simply as a tool fostering community participation or engagement. We provide evidence of the importance of integrating an infrastructural approach to the partial view of social media as networks. We conclude by

discussing the evolution of social media infrastructuring technologies and the making of the "social web".

Introduction

Social media and the technologies that sustain them have dramatically changed how users interact and access content, products and services on the web (Aral et al., 2013). Apart from general social media such as Facebook and Twitter, many industry specific platforms are now being established in almost every business sector. In most of these contexts, social media set the terms of user platform participation by shaping the means through which users interact online. They also influence the modes of data production and exchange on the web, through the innovative use of so-called social technologies. Social buttons, authentication APIs (Application Programming Interfaces), data aggregation and filtering mechanics compose a complex and layered technical infrastructure that is making the web "social". These developments have important repercussions for businesses and organizations. The technical infrastructure of social media conditions the uptake and adoption of services, shapes the success of innovative data-driven enterprises, and dictates new technological advancements across systems and devices.

A few authors have argued that the diffusion of social media coincides with the "platformization" of the web (van Dijck, 2013; Zittrain, 2008), whereby large portions of the internet are shaped by the architectural and operational distinctiveness of social media platforms (Helmond, 2015). Despite all of this, the majority of IS scholarly contributions continues to view social media platforms as predominantly social networking sites (Berger et al., 2014). When focused solely on the social networks they enable, the analysis of social media risks missing what constitutes the novelty and specificity of social media: the

complexity of the technical and social arrangements on the basis of which they produce, circulate and commercialize social data (Alaimo and Kallinikos, 2016). Casting social media in this light may help understanding the role they play in the emerging data economy and the current development of the social web.

Since boyd and Ellison's widely adopted definition of social media as social networking sites (2008), studies on social media have been mainly focused on the role of users and their social network dynamics. The definition contributed to a general "user-centric" perspective that is not well suited to account for and explain the infrastructural dynamics underlying social media and their contribution to the emerging data economy (Brynjolfsson and McAfee, 2014; Varian, 2010). The difficulty in transcending the user-centric approach is closely associated with the fact that social media are mostly *visible* as *social* networks. This is what we here call the social media "visibility effect", whereby social media analysis is bound to what is visible (and oftentimes measurable) as links between users. As a consequence, the underlying socio-technical dynamics of social media and the layered architecture sustaining it remain mostly uncharted. The vivid research interest in networks and the massive amount of data available have enabled the study of social media as networks and, at the same time, limited it to a "user" network analysis at the interface level.

It is of utmost importance, we believe, to integrate the user-interface view of social media as networks with a view that addresses the specificity of social media as socio-technical arrangements of structural depth (e.g. complexity, layering). In this paper, we focus on three interconnected questions that address these concerns. More specifically, we ask: how do the infrastructural conditions underlying social media technologies shape network visibility in the first place? How do social technologies structure the participation of lay publics? What is the role of APIs in filtering social data circulation and commercialization within and across social media networks?

Answering these questions requires first addressing a bigger conceptual challenge as to how social media networks can be better understood if framed as information infrastructures. To this end we critically review the literature on social media and develop a conceptual framework of the generative mechanisms of social media networks. We use this framework to build a model for our empirical analysis. We subsequently use this model to formulate four hypotheses that are tested with a mixed method approach to the study of Last.fm, a social media platform for music discovery (Miles and Huberman, 1994; Yin, 2009). We explain how Last.fm produces and computes social data and how its technological features define what becomes visible at the interface level. We show the strong association of Last.fm social technologies on weekly user listening activity. Our findings suggest that social media are more than networks and that user behavior is more strongly linked to the underlying socio-technical infrastructure than to social participation. The primacy of social media technologies suggests that firms cannot view social media simply as a tool fostering community participation or engagement. We provide evidence of the importance of integrating an infrastructural approach to the partial view of social media as networks.

Our research contributes to the extant IS literature on social media and to the literature on social (media) networks. We call for an integrated research agenda that extends beyond the interface networks and considers the importance of the infrastructural attributes of social media. In doing so, our research builds on and extents the literature on information infrastructures (Hanseth and Lyytinen, 2010; Tilson et al., 2010; Yoo et al., 2010). Our paper further builds on existing research on social media networks (Aral et al., 2013; Kane et al., 2014; Leonardi, 2015; Sundararajan et al., 2013). It claims that IS research and the focus on information infrastructures can greatly improve our theoretical understanding of social media platforms and how they enable or constrain the emergence of social media networks.

The paper is structured as follows. First, we critically review the literature on social media and show its network predilections. We summarize our theoretical understanding of social media as "more than networks" building a framework that conceptualizes some of the key social media technologies as the generative mechanisms of social media networks: the self-reinforcing cycles of social data production, filtering and ordering, the role of APIs as boundary resources, and the role of lay public participation for the production of social data.

Second, we present our empirical case of Last.fm, a social media platform for music discovery. By drawing on our theoretical framework, we focus on Last.fm main technological features and we explain how they operate. Subsequently we show the influence of Last.fm core technological features to user listening activity, which is instead only tenuously linked to community participation. The empirical findings attest to the need of considering the infrastructuring properties of social media technology as the generative mechanisms of social media networks and consequently as major players in the study of user behavior. We conclude by discussing the evolution of social media as infrastructures and their contribution to what we believe to be the making of a "social web".

Literature Review

Are social networks on social media different from social networks offline? Social networks are commonly defined as groups of people interconnected by social interactions and personal relationships. Social network analysis (SNA), the methodology used to study social networks, defines people as 'nodes' and their connections as 'links' (Barabasi, 2002; Borgatti et al., 2009; López and Scott, 2000; Wasserman and Faust, 1994).

Should social media networks be differently studied from social (offline) networks? This is a fundamental question raised by several IS contributions (Berger et al., 2014). In what follows

we review the literature on social media showing a conceptual mismatch between the layered complexity of these platforms and the methodological frame of SNA. We articulate our critical review along the following lines: 1) the analysis of social media as social networks focuses on a "here and now". It offers a local view which overlooks the invisible and distributed socio-technical legacy that oftentimes leads to network visibility and network emergence; 2) the structural analysis of social media as social networks needs to assume the network as predominantly stable (otherwise the links that define the network lose visibility), while social media are inherently dynamic and in constant flux. By the same token, it needs to assume the network as a closed, bounded system, while social media are sustained by continuous and expanding data flows and intra-platform interactions; 3) SNA is based on the assumption that people are the nodes of the network and their relationships are the links. The general approach of SNA is "content agnostic", it doesn't look empirically at what exactly flows in the network and what consequences it may have on the nature of networking activities (Sundararajan et al., 2013). In general we may say that, when SNA is applied to social media, it does not adequately consider its technological specificities. There is little or no technology of social data production and circulation in the social network study of social media. Yet, not only are social media networks formed on top of the technological layers of production, filtering, storing, transmission and feedback of social data, but also they are, first and foremost, also social data networks resulting from the data operations mentioned above.

Beyond the network façade

As remarked earlier, an important step in the analysis of social media has been the definition of social media as social networking sites (boyd and Ellison, 2008). Boyd and Ellison defined social networking sites as: "web-based services that allow *individuals* to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those

made by others within the system" (boyd and Ellison, 2008). Since then, the majority of contributions have more or less explicitly analyzed social media by adopting a social network perspective. Social networking sites were web-services that enabled: "users to make visible their social networks" (boyd and Ellison, 2008, p. 211). As this claim points out, the two concepts that have been pivotal for the study of social media are "user" and (network) "visibility". In fact, social media have been analyzed mostly as visible social networks and tacitly catalogued as "user enabled". Social media have been defined as social networking sites, bounded systems, where IT has been viewed as a neutral facility of user interaction (Lovink and Rasch, 2013). Technology, in other words, has little or no role to play in the way in which user-enabled interaction takes place. The system acts as a neutral space, a "neutrality" that should be constantly questioned, particularly when hidden behind the concept of platform (Gillespie, 2010).

A number of scholarly contributions in business and management studies have started questioning the role of technology in social media network articulation (Aral and Walker, 2011; Oestreicher-Singer and Sundararajan, 2012b; Oestreicher-Singer and Zalmanson, 2013). Kane et al. (2014) recently called for further research on 'social media networks' and their impact on organizations. Their paper stresses the need of moving away from classical SNA when studying social media networks. Drawing from Monge and Contractor (2003), Kane et al. (2014) remark on the newness of a "*social media-enabled networks*". They recognize that these social media networks are *different* from social networks because they have been enabled not just by users, but also by social media technology (that is why they are called social media networks). However Kane et al. (2014) overlook the technological distinctiveness of social media networks and remain focused on their visible features, including: "digital profiles, relational ties, search and privacy, and network transparency." In a similar fashion, Oestreicher-Singer and Zamalson (2013) analyze user willingness to pay on

Last.fm showing its strong association to community participation. Their model does not consider if and how willingness to pay is also associated with Last.fm core technology and core business: its personalization service. Last.fm is primary a music discovery service powered by a recommender system based on user activity. Working similarly to Amazon the collaborative filtering in place ranks and therefore orders the display of information (in this case artists) on the platform: it makes visible both content and community in particular ways. Overall, the majority of contributions on social media networks, even when they analyze social media "visible features", do not explain how they are made visible in the first place (Kane et al., 2014; Leonardi, 2014, 2015; Oestreicher-Singer and Sundararajan, 2012b; Oestreicher-Singer and Zalmanson, 2013).

The lack of studies on network emergence is a long and vexed problem in network analysis (Emirbayer and Goodwin, 1994). Several authors have claimed that SNA doesn't account for network generation or causal mechanisms but, given a set of nodes and ties (or links) and a mathematical model, it just predicts the behavior (power distribution) of the elements of the set considered (topological analysis) (Emirbayer and Goodwin, 1994; Knox et al., 2006; McPherson et al., 2001). Despite various calls for research there is still limited understanding on network formation and evolution (Yan et al., 2015). Relying on social network analysis opens many roads to the study and classification of network dynamics but at the same time avoids confrontation with the "invisibility" of social media technologies: the enablers of network emergence (Probst et al., 2013).

A stream of very recent contributions in the field of media studies points out to the importance of the invisible technology in shaping what becomes visible on the platform interface as network and the extent it shapes how users interact (Bucher, 2012, 2015; Elmer et al., 2015; Gerlitz and Helmond, 2013; van Dijck, 2013). The common approach of this strand of research is to go beyond what appears as network at the interface level, seeking to

account for how technological features intervene in the process of visibility: that is, how specific technological functionalities are deployed to organize and order the display of data and information at the interface level, thus enabling or constraining what users see and how they interact. Beyond the surface of social networking there is a constant technological infrastructuring that takes shape as links between technical components (rather than users) and the data operations they enable (van Dijck, 2013). As Bucher points out, network visibility is first and foremost a product of the specificity of the (technological) medium (Bucher, 2012). Visibility and its possible network articulations, in other words, is the result of a specific social media socio-technical (invisible) infrastructure.

As Oestreicher-Singer and Sundararajan have pointed out, visibility plays an important role in network dynamics altering their socioeconomic impact (Oestreicher-Singer and Sundararajan, 2012b; Sundararajan et al., 2013). The way users see something online (i.e. other user choices, products, comments) conditions the choice they will make. In a different context, Leonardi has demonstrated the effects that enterprise social networking technologies have on work practices in organizations, by making visible something that was not visible before (Leonardi, 2014, 2015).

Understanding how visibility is produced on social media is a prerequisite to study the behavior of networks and their socio-economic impact within and across platforms and organizations. What is it that makes "stuff" visible in the first place? How? If we look at Facebook, it is EdgeRank, the algorithm that orders the personalized display of information on the NewsFeed (Facebook homepage) that dictates platform visibility. As Bucher has demonstrated (2012), users who want to become more visible behave by following the rules of EdgeRank. They try to use keywords or to obtain more "Likes" in order to jump on top of the NewsFeed ordering in a sort of social media "algorithm optimization" (Introna and Nissembaum, 2000) that has little to do with classical social networking dynamics. Needless

to say, if one's profile is not visible it cannot be part of the network (Bucher, 2015). The "rich gets richer" in other words is not just a product of network dynamics, but also the result of algorithmic logic. It is the ranking embedded into the algorithmic ordering that gives relevance and therefore visibility to users on Facebook. The "visible features" that are observed in SNA are themselves the result of complex computational processing and of the infrastructural arrangements that tie algorithms with real-time data. Predicting network behavior without understanding the influence of these computational tools and arrangements gives just a partial view of these phenomena that result from the complex technological and computational devices and automations that lie on the backend of these machines (Elmer et al., 2015; Gehl, 2014; Gerlitz and Helmond, 2013).

Beyond boundaries

Social network analysis studies the structure (who is linked to whom) and the behaviors of the nodes (the consequences of individual actions for the entire networks) as two distinct but interconnected characteristics of the network. The analysis of networks thus derives from developing mathematical models to read the network structure (usually mathematical models and graph theory) and behavioral models to read and predict the behavior of the network (Easley and Kleinberg, 2010; López and Scott, 2000).

In order to perform the structural analysis of social networks, SNA needs to presuppose the network as stable and closed. Already in 1976, White et al. adopted a more cultural approach to network analysis that stressed the need of accounting for the causality of external forces in tracing changes in network configurations overtime (White et al., 1976). As Sundararajan et al. remark, extant research is based on the assumption that networks are more or less static (Sundararajan et al., 2013). When SNA accounts for changes over time it does it through succession of static representations of its structure (Emirbayer and Goodwin, 1994).

Even if we consider the issue solely as methodological, it nonetheless has consequences in the way social media networks are defined and studied. SNA tools allow modeling of structure and content, or the relationship between structure and content, but in order to do so the sample needs to be fixed. A consequence of this is that exogenous forces cannot be taken into consideration (therefore the stasis and closure of the network). These analytical requirements appear to be particularly problematic when one considers the kinds of large-scale, integrated, interconnected and layered technologies that make up the social media ecosystem today (Bucher, 2015; Tilson et al., 2010; Yoo et al., 2010).

It is very difficult, and will be even more difficult in the future, to impose analytical stability and fixed boundaries to systems and platforms that are predominantly based on social data production and exchange (Alaimo and Kallinikos, 2016). To repeat, SNA also presupposes networks as having fixed boundaries; that is, when it analyses social media networks it assumes them as being confined within the boundaries of a technological system (boyd and Ellison, 2008, Ellison and boyd 2013, Kane et al. 2014). One of the most interesting aspects of social media today is exactly the absence of boundaries or better, the way in which social media technologies are deployed in constantly renegotiating boundaries (Eaton et al., 2015). Social media enabled networks rest on a continuous flow of social data across single platforms, applications, devices and business domains. It is this that has been defined as the "social web"; namely, a layered ecosystem of interconnected platforms that is powered by the production and circulation of data and by the technology that makes this circulation possible, such as for instance APIs (Helmond, 2015).

Social data are the data produced by user platform participation. The fact that they are at the core of social media functioning is well exemplified by Facebook. One of its main strengths derives from the Open Graph, its network of data flowing in and out of the platform. Since the inception of Facebook's Open Graph, as Gerlitz and Helmond have demonstrated, a

continuous flow of data is able to generate economies of scale, what they have called the "Like Economy" (Gerlitz and Helmond, 2013). They refer to the diffusion of the "Like" buttons across the web, the distribution of social technologies from Facebook to a "socialized web" that makes possible the production and exchange of "Like" data. With the implementation of the "Like" button users "Facebook-like" digital objects all over the web and the data produced are re-directed back to Facebook where they enter in the platform's broader economic exchanges. Clearly, as Gerlitz and Helmond have pointed out (2013), the social data exchange powered by social buttons and APIs is more than a network, it is a "Like economy". Facebook "Like" is perhaps the most famous case of how social media create value by producing social data "enabled networks". As the example demonstrates, social media data production and exchange are enabled by a distributed infrastructuring technology that has a primary role in making the web social. In fact they seem to mark yet another stage of social media evolution, whereby platforms seem to evolve toward infrastructures: dynamic elements in a web of socio-economic actions that they constrain and enable at the same time (Hanseth and Lyytinen, 2010; Tilson et al., 2010).

The ever-changing nature of social media platforms clearly makes the stability assumption of SNA very problematic and the value of snapshot mappings of network structure questionable. Facebook routinely changes EdgeRank, the algorithm that regulates the visibility of its NewsFeed (user activities on the homepage). EdgeRank now responds to more than 100.000 personalized outputs in regulating NewsFeeds for each and every Facebook user (Alaimo and Kallinikos, 2016). This means that, every time one of the 1.19 billion monthly active users, 874 million mobile users, or 728 million daily users of Facebook[4], does something relevant

4 Statistics are sourced from The Next Web. See: http://thenextweb.com/facebook/2014/01/29/facebook-passes-1-23-billion-monthly-active-users-945-million-mobile-users-757-million-daily-users/ (Last accessed on the 19th of November, 2015)

for the algorithm functioning, within or beyond Facebook, his or her "network" appears - that is, it is made visible by EdgeRank - in a different way.

More than data, less than content

SNA measures how resources flow in the network but rarely relies on empirical evidence to demonstrate how the nature of these resources impacts their propagation in different networks. The majority of theories in sociology, marketing, economics, and IS, from Granovetter's strength of weak ties (1973) to Burt's brokerage effect (1992) rely on assumptions on the nature of resources flowing. This approach, which Sundararajan et al. (2013) call "content agnostic", avoids investigating the very nature of the resources flowing within the networks to validate or refute assumptions. Seldom does SNA focus on the "nature" of the nodes or ties, it instead 'classifies' them on the basis of existing typologies (either drawn from the network environment or from the behavior of particular network configurations) (Zeng and Wei, 2013). Because SNA presupposes the nature of the entities linked as well as their intrinsic qualities, it does not give an account of whether and how those nodes or entities get transformed by being linked or propagated in the network. "Network theory builds its explanations from patterns of relations", Ronald Burt observed, "bypassing (…) the significant attributes of people" (1986, p. 106, quoted in Emirbayer and Goodwin, 1994). Network analysis does not consider whether the network is made of people, data, or "stuff" and cannot explain how those entities get transformed by being propagated (Emirbayer and Goodwin, 1994).

One of these assumptions regards the production of content. When studying user generated content (UGC), typically network approach contributions focus on user participation rates or on the motivational factors behind user participation (Zeng and Wei, 2013). Studies on social media networks usually take for granted the very concept and nature of what is "content", how its production is enabled by specific technologies for specific purposes, and how the

recursive relation between user participation-technology-content is constantly redefined at every user click. The reliance on an unpacked notion of content has conditioned the very definition of different social media networks. Social media networks have been defined by looking at *typologies of content* flowing through the network. For instance, Berger et al., following (Beer, 2008), distinguish "user-oriented sites" in which networking is the main preoccupation" (p. 518) from "content-oriented sites". Among the latter they list sites such as Twitter, YouTube, and Flickr that have inherited some features of what they call OSN (online social network) but are more focused on "blogging" activities, that is, the production of content by users (UGC) (Berger et al., 2014). Classifying social media on the basis of content typologies overlooks the nature of content, its technologies of production, and the ways the two are bound together by specific platform configurations (Kallinikos et al., 2013). Is Twitter user-oriented or content-oriented? Shi et al. for instance, see Twitter as a social broadcasting technology (Shi et al., 2013).

Other scholars have similarly defined social media as "technologically enabled networks" distinguishing between networks that mostly connect (or facilitate connections between) users, and networks that facilitate the flow of products (or content such as for instance music as in the case of Last.fm, the empirical object of this study). This is the position of Sundararajan et al. who define "digital networks" as IT artifacts created to facilitate interaction, arguing that they are of different kinds (Sundararajan et al., 2013). Meanwhile networks of "tags", (such as delicious.com), have been studied as defying traditional categorization as communities or social networks (Levina and Arriaga, 2014). By contrast, something like Facebook instead is seen as a *digital approximation* of (well-understood, familiar) social networks (Sundararajan et al., 2013, p. 885, italic added). A definition of this sort risks overlooking important theoretical insights on the emergence of new network configurations as in the instance of the "Like Economy" previously mentioned (Gerlitz and

Helmond, 2013). Is a "Facebook Like" content? What is the relation between "Like buttons", user participation, and "Like data"?

The critical point here is that social media networks are digital networks and their digitality needs to be unpacked. Being technologically enabled, social media networks defy existing classifications of content or resources flowing in the network and call for a more precise definition of what the stuff is that flows in the network and what its socio-technical emergent properties are (Kane et al., 2014). Starting from the useful reminder that online everything is data, it is necessary to unpack the notion of "content" and the way it is bound to the participation of lay public for the production of social data. To what extent content does production responds to specific social technologies? How does it structure (is structured by) user participation?

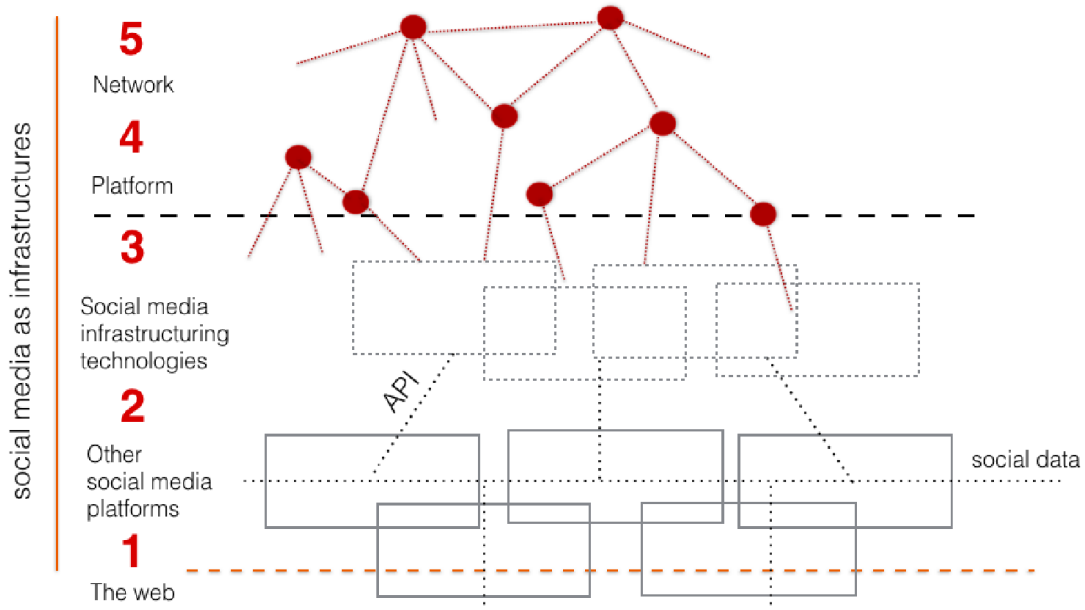
The generative mechanisms of social media networks: social media as infrastructures

The study of social media requires more that a network approach and more than an analysis of information transmission. We contend that framing social media as information infrastructures could help theorizing the role of IT in structuring and articulating the very social media network under study and its transformative effects on user behaviors.

Henfrindsson and Bygstad recently asserted that "[t]he notion of infrastructure has been adopted as a way of conceptualizing interconnected systems collectives (rather than stand-alone information systems)" (Henfridsson and Bygstad, 2013, p. 908). Social media are different from stand-alone systems, as information infrastructures they depend on pre-existing infrastructures and constantly evolve over time. They are "shared, evolving, open, standardized, and heterogeneous installed base" (Hanseth, 2002).

x t a i e f ng social media as infrastructures. On the one hand, s g level 5 in figure 1) rest on specific platform features (what vi i on level 4 in figure t i m a e o i t e erally belong to the y e i a i anking: what make the visible f e i n figure 1). On the ha s a t c o i echnologies: social buttons, social bookmarking, APIs, technologies that plug- cific social media t in figure 1). Social i a s r i t tforms, and devices which themselves rest on the Internet as complex infrastructure seth and Lyytinen, 2010).

Figure 1: Social media as infrastructures.



nne w s ne be be e cated and theorized (Henfridsson and Bygstad, 2013; Tilson et al., 2010). o pose to look at the e og a a ve of ial media networks

leading to specific network emergence and configurations (Archer and Bhaskar, 1998; Henfridsson and Bygstad, 2013). Theorizing the missing links between the invisible infrastructural properties of social media and precise socio-technological features as the generative mechanisms of social media networks may offer a better understanding of the complex role of social media in today's data economy. Generative mechanisms are defined as the causal structures that generate observable events. Generative mechanisms are unobservable and non-deterministic, however their effects can be observed, constituting a good explanation of how the mechanism works which in turn becomes generalizable (B. Bygstad et al., 2016).

Observable events in the case of social media are for instance, the "visible features" commonly studied: profiles, commenting, posting or sharing functionalities, etc. (level 4 in figure 1). The "visible features" are generated by elaborate information infrastructures and the computations they enable. We want to study three of generative mechanisms leading to social media visible features.

The first is the self-reinforcing process of data production, organization and display at the interface level, which is generated by invisible mechanisms and actuated by the contingent users interaction. These data flows are sustained by constant technological adjustments to user participation (and vice versa) (Ciborra, 2000). A good example of this self-reinforcing generative mechanism of social data production and ordering leading to visibility is the interplay between Facebook's EdgeRank and user participation on Facebook NewFeed as previously mentioned (Bucher, 2012). Capturing how this self-reinforcing cycle of social data production happens is a necessary condition for understanding network emergence and dynamics. As known, the visibility of these networks alters their socio-economic impact.

As seen though, social media are large scale, distributed, interconnected and layered infrastructures. The production, filtering, and ordering of data does not happen within the

boundaries of a system but within and across platforms, applications and devices. Therefore, the second set of generative mechanisms we want to look at are APIs as socio-technical boundary resources. A boundary resource enables coordination of activities and data across multiple socio-technical worlds where heterogeneous but co-operating actors are tied together (Eaton et al., 2015; Ghazawneh and Henfridsson, 2010; Star and Griesemer, 1989). On the one hand, a boundary resource is what opens the design of the platform to third party developers allowing the proliferation of applications built on top of social media infrastructures. On the other hand, a boundary resource is what maintains a certain degree of control by purposely excluding particular design choices or data types or even devices. APIs and other boundary resources therefore enter in shaping network configurations. The Facebook "Like economy" with its set of APIs is the observable result of these type of generative mechanisms (Gerlitz and Helmond, 2013).

Finally, one of the most interesting generative mechanisms of social media is how they structure the participation of lay publics. Much has been written on the issue of participation on social media. For the purpose of this paper we follow Zittrain to critically question the role of lay public participation in relation to content and social data production (2008). Zittrain saw the generativity of a system as its capacity of producing unanticipated change fueled by the participation of broad and varied audiences (Zittrain, 2008). So far the participation of lay public has been just related to the generation of content. As seen from the literature reviewed however, technologies of content production have been rarely investigated. Very little is known of how new forms of digital content are tied to novel user behavior or emergent system functionalities (Kallinikos et al., 2013). To what extent is the participation of the lay public not just an engine of data for social media functioning and their connected data economies? Social media defy traditional classification of network content (Alaimo, 2014; Alaimo and Kallinikos, 2016). The generative and transformative effects of digitized content

production by lay publics may produce unanticipated changes under the form of data, for instance that may be fed back to the system and create unanticipated changes.

Table 1 below summarizes the main points we have isolated from the literature and connects them with the conceptual framework. The first two columns synthetize the shortcomings of the network approach and consequent need of framing social media as infrastructures. The third column connects the literature so reviewed and the questions we have isolated from the literature to the conceptual framework: the generative mechanisms of social media networks. Drawing from it, the fourth column spells out the research questions and corresponding hypotheses that have guided the empirical analysis of Last.fm.

Literature review: (More than) Social networks	The need to theorize on social media as infrastructures	Conceptual Framework: the generative mechanisms of social media networks	Deriving a conceptual model for a multi method research design of Last.fm
(1) The analysis of social media as social networks focuses on the "here and now". It offers a local view of networks, which overlooks the underlying socio-technical legacy that oftentimes shapes network ties and leads to network emergence. Lack of studies on the role of technology in enabling or constraining network emergence and visibility.	The "visible features" that are observable in SNA are the result of elaborate information infrastructures and the computations they enable. Social network visibility is first and foremost a product of the specificity of the (technological) medium. The process of visibility can be defined as the organization and display of data and information at the interface level performed by a battery of underlying technologies. These data flows are sustained by constant technological adjustments to user participation (and vice versa). Capturing how it happens is a necessary condition for understanding network emergence and dynamics. As known, the visibility of these networks alters their socio-economic impact.	Self-reinforcing mechanism of social data production and ordering (ranking)	Research question (1) How do the infrastructural conditions underlying social media technologies shape network visibility in the first place? Hypotheses to be tested: Does network visibility impact user listening activity demand on Last.fm? [H1] How does technology intervene in the process of visibility? [H1a; H1b]
(2) The structural analysis of social media as social networks needs to assume the network as predominantly stable, meanwhile, social media are inherently dynamic. Equally, it needs to assume the network as a closed, bounded system meanwhile social media are sustained by complex patterns of endogenous and exogenous data flows.	Social media are large scale, interconnected and layered infrastructures. Social media are essentially based on social data production and exchange within and across platforms, applications and devices. A network of data facilitated by APIs and other socio-technical boundary resources make those systems semi-open.	Boundary making	Research question (3) What is the role of APIs in filtering social data circulation and commercialization within and across social media networks? Hypotheses to be tested: Does user participation impact user listening activity? [H3] Does across platform content availability impact user listening activity? [H4]
(3) Defining and analyzing social media as networks assumes both the nature of nodes (for instance social networks usually conceives people as nodes of the network) and links. The general approach of SNA is "content agnostic", it doesn't look empirically the nature of the flows in the network and the consequences such flows have. On social media every action-link transforms the nodes involved in the exchange and the entire network as well (in real time).	Social media defy traditional classification of network content. The links of the network or the resources flowing in the network call for a more precise definition of what is content (for instance music), what is user generated content (for instance "tags") and what is data. Furthermore the transformative effects of digitized content circulation need to be better analyzed and acknowledged.	(Infra)structuring of lay publics participation	Research question (2) What is the role of user participation for social data production? Hypotheses to be tested: Does UCG impact on user listening activity? [H2] Do different kinds of UGC (i.e. text, image) have different impacts on user listening activity? [H2a, H2b]

Table 1: Exhibit connecting the literature reviewed (column 1 and 2) with conceptual framework (column 3) and conceptual model (column 4).

Given this theoretical background, our empirical study design adopts a theory driven multi method approach (Mingers, 2001). For the qualitative part, we gathered and analyzed data from secondary sources: among which tech blogs, Last.fm website and Last.fm official blog. In particular we have systematically analyzed all the Last.fm blog entries since the beginning of the blog activities (2007). Since its inception the blog has been written and curated by Last.fm staff and even if it has been recently removed from Last.fm we have retained copies of all the material analyzed. Data have been selected on the basis of the literature reviewed. Data selection and analysis have been theory driven with the aim of constructing a retroductive explanation of possible generative mechanisms of social media networks (Archer and Bhaskar, 1998). We wanted to understand Last.fm functioning as infrastructure, to which end we analyzed the data gathered and we constructed a schema of Last.fm architecture (see figure 2). The schema helped us visualizing Last.fm as layered architecture and identifying three generative mechanisms: the self-recursive cycles of data production, the system of APIs and the role of UGC or user participation for the platform functioning (Miles and Huberman, 1994; Yin, 2009).

The quantitative part of our design is a confirmatory study of the generative mechanisms of Last.fm. It tests the robustness of our hypotheses on the relevance of social technologies to network emergence and user behavior. We provide a model that shows the strong association of social technologies to social media networks and user listening activities (more details in what follows). The quantitative phase data collection and methodology are spelled out in the following section. In the next session we first introduce Last.fm, the empirical object of our investigation and its functioning as infrastructure of music discovery and consumption.

Empirical study: Last.fm

Last.fm is currently one of the oldest and has been one of the most popular online music discovery services. It was established back in 2002 and was acquired by CBS for $280 million in 2007. Since its inception to early 2009, users could stream free music directly from the service. In April 2009, the company limited free streaming to the US, UK and Germany, citing an inability to recover music licensing fees from advertising[5], meanwhile users in other countries were required to pay a subscription fee. Over the last five years, Last.fm has gradually wound down all streaming operations to focus its business exclusively on music discovery. Last.fm operates with 'freemium' business model, whereby basic services are provided for free, and premium services are offered for a fee, which together with advertising constitutes the main revenue sources of the company.

Last.fm core activity is to suggest music to its user base by collecting and computing data on user listening behavior taken from partners such as Spotify[6] and YouTube. Last.fm powers music discovery service with its proprietary technology called 'AudioScrobbler'. This is essentially an item-based collaborative recommender system (Ekstrand et al., 2010; Jannach et al., 2010; Konstan and Riedl, 2012; Riedl and Smyth, 2011) that works by computing data on listening behaviors 'playback events', collected through Application Programming Interfaces (APIs) from more than 600 playback applications, services and devices distributed across the web.

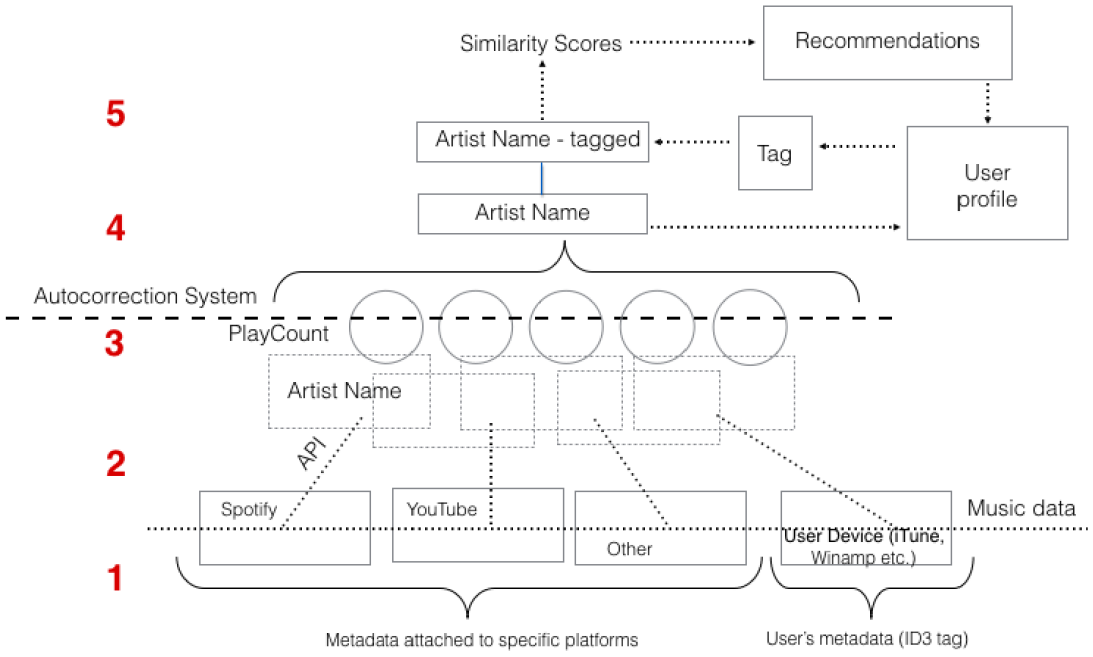
Figure 2 illustrates the role of APIs as boundary resources connecting Last.fm to external devices, application and platforms and allowing data on user listening behaviors 'artist names' to be ingested by the Last.fm system. The system will then 'count' the 'artist names'

5 From Last.fm blog: http://blog.last.fm/2009/04/22/radio-subscriptions

6 From Last.fm blog: http://blog.last.fm/2011/11/30/lastfm-for-spotify

e t i ‘ (or ure g). ‘Play Count’ or y e he rt ’ s i m from APIs. More precisely, users that download ‘AudioScrobbler’ or related plug-ins and Last.fm API automatically would be able to submit user listening data fro ected 600 playback d vi s v out c unt’ is the result of s h s ide Last.fm). ‘Play count’ is the data entity at the ba m t ng recommendation system.

Figure 2: Last.fm layered architecture



Item- ol bo ve similarity ranking. f ’ s lations by mapping b ‹ks’ on the basis of i ri l 2, I ıst.fm computes the probability that users t o a e t o s nilar artists.

Last.fm relies also on social media "visible" features that are characteristic of ordinary social networking sites. For instance, users have profile pages, which display user activities as 'playback events'; they can add other users as their friends, and participate in the social communities in various ways: creating and leading groups, or joining existing groups, where they can start to participate to discussion threads on things of their interest. Users can also chose to contribute to the platform in more substantial forms; for instance by writing blogs about music, writing biography (wiki) about artists, or uploading images, music or videos of their favorite artists. Users can also add information regarding music events and invite other users or join upcoming music events.

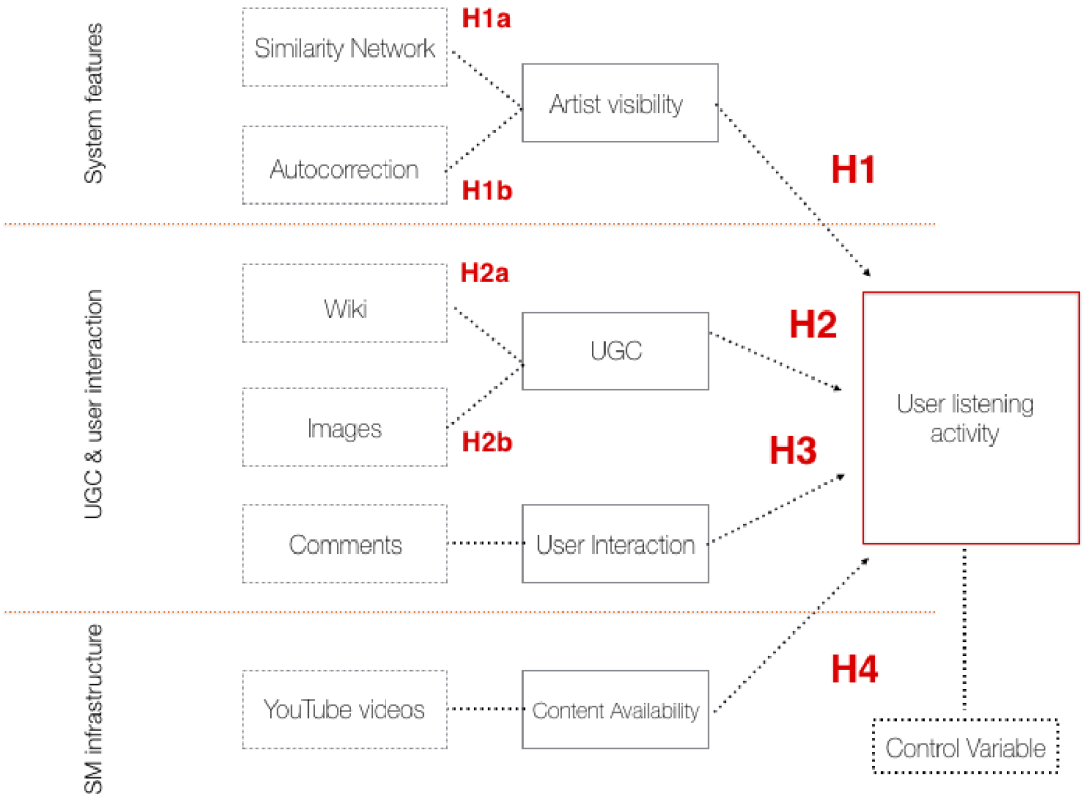
One of the most important activity of users (beside producing listening events) is to 'tag' artists, albums, and tracks with any keywords. In fact 'tags' enter into the 'Audioscrobble' of Last.fm to construct 'similarity networks'. The reason is simple: computing similarity scores solely on the basis of user listening behavior may be problematic. For example, users who like to listen to classical music may also like to listen to rock music. A score based solely on listening data would determine classical and rock music as similar to one another. To attenuate this problem, artists on Last.fm are deemed similar to one another not only when they are being listened by the same group of users, but also when they are being labeled with the same 'tag'.

Conceptual Model

Figure 3 below illustrates our conceptual model of user listening activity on social media for music discovery. It operationalizes our framework, the generative mechanisms of social media as variables. It spells out the association of each specific social technology (data, production cycles, API and user participation) to user listening activity (number of weekly

t he t i e m s ructures by showing the strong association e ne c and to user behavior (in this case operationalized as weekly user listening activity Be e delving into the l takes into account r l r w re ology operate. For a y l l c sy em features of data production cycles. These are the key *system* a t ollaborative filtering s e . The middle layer accounts for user participation and interaction thus representing social *pl tform* features. The third layer instead accounts for the social web *infrastructuring* c ogy, which allows tom end).

Figure 3: Conceptual Model



Conceptual model variables

H1 ‘Artist visibility’

As mentioned, the process of visibility refers to the ways in which specific technologies organize and order the display of data and information at the interface level. Visibility is an essential aspect of social media and social media networks functioning that it is usually taken for granted. Here we analyze Last.fm ‘artist visibility’. On Last.fm ‘artist visibility’ is conditioned by two fundamental operations: ‘similarity network’ and ‘data aggregation and autocorrection’. It is because of the complex machinery behind these two operations that Last.fm is able to construct a ranking of popular artists that is displayed on the website and therefore made visible under criteria of relevance for users. Our H1 is thus:

> H1: Higher artist visibility is positively associated to higher user listening activity (i.e. consumption of music from the artist)

H1a ‘Similarity network’

‘Similarity network’ is the network of similar ‘artist name’ the main data entity of Last.fm. It is essential to understand that on social media networks there is an intrinsic equivalence between “social networks” and “content networks”. Because of the computational processes in place, they are interchangeable, flipsides of the same coin (Jannach et al., 2010; Seaver, 2012). On Last.fm ‘similarity network’ forms a crucial component in the operation of ‘Audioscrobbler’ to produce personalized music recommendations. It is because of the sustained network of similar artists that Last.fm computes the degree of similarity between different artists. The criteria by which artists are deemed similar do not only construct personalized suggestions; they also shape the organization of information made available to users. When users browse to discover content they actually navigate the Last.fm ‘similarity network’. As discussed above, ‘similarity network’ is constructed by listening data and ‘tag’

data. Artists are more visible if they obtain a high similarity score as artists are being ranked according to this score on web pages for users to navigate and users are more likely to pay to attention to items ranked higher.

Last.fm can only assemble list of similar artists for artists with more than 5 listeners. Those artists without lists of similar artists will not be recommended at all by the recommender system of Last.fm since similarity network forms a crucial ingredient of the production of personalized recommendation.

> H1a: Relatively high similarity score is positively associated with higher artist visibility

H1b 'Autocorrection system'

In order to set a similarity network in place, Last.fm performs different operations of data collection, aggregation and filtering. As mentioned, Last.fm does not stream music; it rather collects 'artist names' data by counting 'playback events' (user listening behavior data) from more than 600 playback applications, services and devices. This means that there is no content flowing in the system (if by content we mean 'music') but just data on user listening behavior, that is, 'playback events'. Similarly to all other social media, Last.fm developed a system of APIs that enable the flow of data in and out of the system, allowing third party developers to establish connections between their playback applications, services and devices with Last.fm. As mentioned, APIs as socio-technical boundary resources (Ghazawneh and Henfridsson, 2010) perform as bundles of functionalities and standardized procedures that make possible the flow of data between different platforms or systems, here the submission of 'playback events' (Yoo et al., 2010). To flow smoothly within and across systems data need to conform to the requirements set out by boundary resources such as APIs.

'Track.scrobble' the method for adding 'playback events' to a user profile of Last.fm, for instance, requires at least the three parameters of artist names, track names, and timestamps for 'playback events' to be successfully submitted. Furthermore, 'playback events' are ingested only when the track is longer than 30 seconds and has been played for at least half its duration (or for 4 minutes, whichever occurs earlier). Because of the aforementioned filtering criteria, specific genres of music such as *grindcore* that have tracks of short duration are automatically excluded. The same filtering criteria impact also on tracks of long duration whose frequency of 'playback events' is likely to be less[7]. However these filters represent just the beginning of very complex data cleansing operations, even after these procedures, the submitted 'playback events' go through numerous other filters[8].

The lack of suitable institutionally controlled music identifiers further condition the music industry data infrastructure. There is no agreed upon standard to 'name' digitized music. Because of this, each system uses its own standard. Last.fm uses artist names, which have to be submitted with 'playback events' as pigeonholes to qualify data. Despite the filters in place, Last.fm needs to implement a set of procedures to correct data inconsistencies caused by the music data-legacy. To this end, Last.fm has implemented 'Autocorrection' in January 2009, which identifies incorrect 'artist names'[9] and points or maps them to their correct

---

7 To solve this problem some users of Last fm have suggested a quantification of 'playback events' according to duration of the music played instead that 'play-count' (see Last.fm forum: http://www.last fm/forum/21717/ /623771/1).

8 Among them is artist name filter, and artist names, which Last fm filters out, includes 'Unknown Artist' and those that appear similar to filename. Besides, Last.fm discourages third party developers from determining metadata from filenames of music and encourages them to use metadata from well-structured sources such as ID3 tag instead. Because of this metadata that appears like filename will be ignored (filtered out) by Last fm's system. This somewhat can cause problem as there actually exist genuine artists with those names, for example, MOSAIC.WAV, which is a Japanese Moe-pop band (see http://www.last.fm/forum/21713/ /70366

9 To identify a correct 'artist name' Last fm follows Musicbrainz convention, which respects the intention of artist (Principle of Artist Intent). See https://musicbrainz.org/. Some of its rules read as follows: correct artist names must have exact spellings used on releases by that artist, if artist names are not consistent across releases, correct names are the most commonly used name, if an artist deliberately uses different aliases for different releases, those names are not mapped to one another (etc.).

versions. For example, 'The Arcade Fire' would be identified as incorrect and mapped to 'Arcade Fire'. It should be clear by now how 'autocorrection' impacts on 'artist visibility'; that is, the system will compute 'similarity network' for 'Arcade Fire', but not for 'The Arcade Fire'. By the same token, trying to access 'The Arcade Fire' would lead to be automatically redirected to the artist page of 'Arcade Fire'. This 'mapping' operation is further complicated by all the metadata associated with incorrect 'artist names' as well as any computations attached to it (for instance, measures or scores of similarity or popularity). That too, should be mapped to correct 'artist names'[10]. Cleaning up the database of Last.fm is a never-ending task as new 'artist names' are constantly ingested into the system[11]. The variable 'autocorrection system' tests to what extent those complex operations of ingesting, cleansing and mapping data impact on artist visibility and therefore on user listening activity.

> H1b: Higher number of associated incorrect names is positively associated with higher artist visibility

H2 'User generation of content' (UGC)

After 'playback events', data is successfully submitted to Last.fm and goes through filters whereby the system updates the corresponding artist pages. If the artist is new to Last.fm database, the system creates new artist pages that need to be filled in with data. It is at this

10 The extent to which this task is being accomplished is still unclear as it is not being discussed in any documentation provided by Last fm.

11 Henceforth, Last fm's autocorrection system therefore needs to be constantly updated. Predominantly, this entails collection of audio fingerprints from users, who play music from their own audio files. These audio fingerprints, once collected and combined, allow Last fm to identify variety of music metadata associated with tracks of music, which generate those audio fingerprints. Nonetheless, autocorrection system, which sets out to correct metadata, can appear incomplete or incorrect itself – some of the mapping can be problematic or it may leave out some artist names that ought to be assigned as incorrect. In these cases, users may suggest Last.fm for correction to be made to its autocorrection system. If enough users make the same suggestions then autocorrection system of Last fm may correct itself accordingly. Despite all the effort, it is important to note that the current autocorrection system of Last fm still leaves some problems unresolved, for instance it cannot disambiguate artists with the same name. If artists happen to have the same name, they will have to share one profile page on Last.fm. Also, it would not be possible to merge different profiles of the same artists into one.

point that user are given the opportunity to generate content. Apart from 'tagging' artists, users may add comments at the bottom of artist pages, add video and photos of artists or write biography of artists (with wiki technology) and blogs about artists. Furthermore the user may vote their favorite artist and even add upcoming events (concerts) and check-in to the event.

Content-enriched metadata in bibliography records are helpful to library users as they try to discover relevant materials (Tosaka and Weng, 2011). Here, content-enriched metadata go beyond basic metadata (such as titles) to include content notes, summaries, tables of contents, sample text, and other publication related material. Artist profile pages on Last.fm are in many essential respects blank templates to be filled by user generated content. The presence of two different types of digital objects (text and image) allow us to test differences in user listening activity due to the presence of different type of content.

H2: Higher quantity of metadata is positively associated with higher user listening activity

H2a: Higher number of wiki biographies is positively associated with user listening activity

H2b: Higher number of images uploaded is positively associated with higher user listening activity

H3 'User interaction'

User interaction and participation play a vital role on every social media platform. Similarly to other social media, on Last.fm users are not meant to be mere receivers of recommendations but active participants in shaping each other's taste and enriching the overall experience of the platform. In a recent study on data based music discovery, Charoenpanich and Aaltonen (2015) found out that people may listen to the same music because it allows them to discuss common music experiences together; having more friends

on social media is associated with larger amount of music listening. It is known that the participation of users does influence peer behavior. Discussion seems to have a primary role for music consumption. We disentangle user production of content from discussion or mere commenting to show possible differences in the association of user participation to user listening activity. We expect that

> H3: Higher quantity of comment (i.e. number of posts from users) is positively associated to higher user listening activity

H4 'Across platform content availability'

The majority of social media platforms are based on layered technologies, which are dynamically integrated in larger ecosystems of platforms and applications. Last.fm is a case in point. Currently users do no longer stream music directly from Last.fm as they are encouraged to stream music via external platforms, such as Spotify and YouTube. The organization of the platform activity as it is, represents an interesting point of debate: can Last.fm still be called or classified as a "content-oriented site"? Digital technology has allowed content discovery services to be loosely decoupled from content streaming services, enabling Last.fm to focus solely on the former with its socially enriched recommender system. Therefore "content" for the platform is now a matter of connecting its database with music files elsewhere. This not an easy task, especially with constantly shifting digital ecosystems (Kallinikos et al., 2013) such as that exemplified by music where the location of music files on the Internet may be unstable and may change all the time without prior notice. We expect that

> H4: Higher content availability is positively associated with higher user listening activity

Data collection and methodology

We retrieve a sample of the most visible artists of Last.fm with APIs. This includes the top 1,000 most popular artists displayed during the week of 3rd November 2014 to 9th November 2014. This is the maximum length of most popular artist chart retrievable via APIs. Then, we retrieve the top 50 most similar artists (similarity network) associated with each of those 1,000 popular artists, culminating to a sample of size of 10,945 artists. Because we retrieve top 50 most similar artists associated with each of those 1,000 popular artists, the sample could have been as high as 50,000 (50 x 1,000) artists. But the sample size turns out to be just 10,945 artists because there is much overlap between names of similar artists. The number of weekly listeners of each of those 10,945 artists during the period of 3rd November 2014 to 9th November 2014 is retrieved and is the dependent variable of this model. Our observational period is 3rd November 2014 to 9th November 2014.

Our model has five independent variables. (H1, H1a, H1b) 'Artist visibility' which is constructed by 'Similarity network' and 'Autocorrection'. To quantify 'Artist visibility' on 'similarity network', we mainly rely upon equation specification as proposed by Oestreicher-Singer and Sundararajan (2012b) with some minor adjustments[12]. We also adjust for the number of incoming links originated from 'similar network' of the top 1,000 most popular artists to our sample of 10,945 artists, whereby only top 10 similar artists are taken into consideration into our model, as they are the most visible. We also adjust for the relevance of the incoming links; we do not merely sum them up but we weight them with the number of weekly listeners of the 1,000 most popular artists from which those incoming links originate. The second variable entering into 'Artist visibility' definition is 'Autocorrection'. It is the

12 The estimation method of Oestreicher-Singer and Sundararajan (2012b) allows us to derive pure impact of visible network while controlling for correlation, which may still occur without visible network. Basically, equation estimated by them include (1) items connected by visible network, and (2) items connected by visible network plus items without connection by visible network, which are similar to those items connected to visible network, as explanatory variables.

number of incorrect artist names associated with each artist in our sample of 10,945 artists. This is the number of incorrect artist names the system anticipates. The 'Autocorrection variable' is related to visibility because users will be directed to pages of artists with correct names whenever they try to access pages of artists whose incorrect spelling is being anticipated by the Autocorrection system. Therefore, the more associated incorrect artist names, the more visible the artist, as there exists more paths whereby pages of artists can be accessed from. Autocorrection is merely mapping from incorrect artist names to a correct artist name. Without it, artists who are misspelled will be lost as users would not be able to access them if they type their names incorrectly. When the autocorrection system anticipates incorrect artist names in advance, the loss will not happen. Also, it is likely for incorrect names to be converted to those of correct names on user profile pages, which users may choose to browse through. To take a glimpse into Autocorrection mapping applied to 'playback events' submitted by users, we select 12,839 Last.fm users randomly and retrieve Autocorrection mapping applied to their 'playback events' in June 2014 (Charoenpanich and Aaltonen, 2015).

(H2, H2a, H2b) The second independent variable for our user listening activity equation is 'User participation'. The two variables considered under this category are the number of versions of artist wiki biographies and the number of artist images uploaded for each of the artists in our sample (10,945 artists). We expect that the quantity of images uploaded or the number of versions of wiki biographies condition their user listening activity.

(H3) The third independent variable for our user listening activity equation is user interaction represented by quantity of user comments on artist profiles. We expect Last.fm users to listen more to artists who have more comments posted on their profile pages as this can arguably be taken as the sign of undergoing user interaction, thus of a social driven experience of music consumption.

(H4) The fourth independent variable in our user listening activity equation is content availability. We expect content availability - operationalized as the quantity of YouTube music videos attached to artist profile page - to condition user listening activity.

(H5) The fifth independent variable in our user listening activity equation is the control variable. A control variable is added roughly in accordance with Oestreicher-Singer and Sundararajan (2012b) to ensure that the coefficient associated with network visibility does reflect what it is supposed to, not a mere demand correlation which can be detected even without the network visibility on product complementary effect[13]. Here the control variable is the number of incoming links from the artist 'similarity network', whereby the whole top 50 similar artists are taken into calculation, weighted by number of listeners of the 1,000 most popular artists, from whom those incoming links originate.

Descriptive statistics

Table 2 reveals the descriptive statistics of both dependent and independent variables associated with our sample of 10,945 in size of our conceptual model. One characteristic of these variables is that there exists a large discrepancy between their means and their medians. For example, while the mean of weekly demand stands at 2,268, the median of weekly demand stands at only 717. This indicates that it is likely for the distribution of these variables to be highly skewed. Therefore, these variables ought to be transformed logarithmically in subsequent analysis[14].

13 Control variable entails both items connected by visible network and items without connection by visible network, which are similar to those items connected to visible network. (See footnote 12 for more details).

14 Transformation takes the form ln (variable + 1). This transformation is essential to shift the shape of skewed distribution toward normal distribution.

Table 2: Descriptive statistics

Variables	Mean	Median	Std. deviation
Weekly demand	2,268	717	5,229
Network visibility	14,674	0	38,654
Associated incorrect names according to autocorrection system	1	0	4
Version of wiki biography	15	9	22
Images uploaded	64	23	220
Quantity of comment	1,006	157	5,183
Content availability	58	41	54
Control variable	67,033	20,539	141,389

Estimation results

After all of our variables are logarithmically transformed, the model of user listening activity (number of weekly listeners) is estimated with ordinary least square estimation, and the result is reported below in Figure 3. The results provide support for the claim that social media like Last.fm are more than networks. Social media are infrastructures and social technology is significantly associated with user listening activity.

(H1) While the coefficient for artist network visibility is statistically significant at 1% level, the magnitude of the coefficient is very low at 0.006. This implies that an improvement of network visibility by 1% would increase weekly demand by only 0.006%. Interestingly, 'artist visibility' is more dependent on 'autocorrection system' than 'similarity network'. In fact, the coefficient for number of associated incorrect names in accordance with 'autocorrection system' is also statistically significant at 1% level and its magnitude is as high as 0.211. This implies that an increase in the number of associated incorrect names mapped by autocorrection system by 1% would raise weekly demand by as high as 0.211%.

Indeed, the more associated incorrect artist names, the more visible the artists, as there exist more paths whereby pages of artists can be accessed from. Interestingly, both coefficients associated with UGC (i.e. version of wiki biography and number of uploaded image) are statistically insignificant.

One important factor to be taken into consideration is the obvious impossibility of quantifying the quality dimension of metadata. This is a factor that can be decisive for our result but which cannot be quantified for regression analysis. The last two variables, user interaction as quantity of comments and content availability, are both statistically significant at 1% level with expected signs. Magnitudes of their coefficients are also relatively large. An increase in content availability and quantity of comments by 1% will increase weekly demand by 0.457% and 0.119%, respectively.

Figure 3: Estimation results

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
(Constant)		-1.190	.086		-13.907	.000
Artist visibility	Network visibility	.006	.002	.019	2.909	.004
	Associated incorrect names	.211	.014	.092	14.951	.000
Quantity of metadata	Version of wiki biography	-.022	.015	-.015	-1.498	.134
	Uploaded images	-.015	.013	-.013	-1.199	.231
Quantity of comment	Comment attached to profile pages	.119	.010	.154	11.485	.000
Content availability	Music videos attached	.457	.010	.362	44.280	0.000
Control variable[b]		.544	.010	.431	56.169	0.000

a. Dependent Variable: user listening activity

b. Control variable according to Oestreicher-Singer and Sundararajan (2012)

c. all variables are in logarithmically transformed

Discussion

We argue that theorizing social media as infrastructures is necessary to understand the role of social technologies for the emergence and dynamics of social media networks and the shaping of the social web. We conceptualize three interlocking generative mechanisms of

social media: 1) the self-reinforcing cycles of social data production and ordering, 2) the iterative operations of boundary making of boundary resources such as APIs, 3) the participation of lay publics as it is recursively structured by social technologies. Drawing from this conceptual framework we conduct a multi method empirical investigation. The first part of our empirical study isolates and analyzes key social technologies of Last.fm as the generative mechanisms behind what become visible on the interface level as networks and network dynamics. The second part of our research measures the association of these generative mechanisms to user behavior. Our model supports our framing of social media as infrastructures, it shows that the production and ranking of social data and the filtering of data and content of APIs have the strongest associations with user listening activity on Last.fm. We find that the underlying infrastructural capacity of social media is strongly linked with what has been narrowly considered so far just as the result of individual user behavior or network dynamics.

In what follows we discuss our model results against the three interlocking generative mechanisms we individuate.

The self-reinforcing cycles of social data production and ordering

Generative mechanisms are defined as the causal structures that generate observable events. "Visible features" of social media such as profiles or social features, user behavior, network structure or dynamics have rarely been explained in terms of their emergence or visibility. We theorized our first generative mechanism as the self-reinforcing process of data production, organization and display tied to the infrastructural conditions and the computation they enable. We conjectured that these invisible mechanisms are strongly linked to user participation.

We provide unique evidence, which establishes the connection between infrastructural conditions of data generation, ranking and ordering and user behavior. On Last.fm 'similarity network' and 'data aggregation and autocorrection' define visibility of artists and are substantially associated with user listening activity. As for our initial Facebook example where EdgeRank defines rank and order of Facebook's NewsFeed, on Last.fm 'similarity network' and 'autocorrection' define rank and order of data artists, therefore conditioning their visibility and popularity. These specific technologies organize and order the display of data they themselves produce (under the form of 'playcount'), thus enabling or constraining what users see and how they interact. As seen, the system does not collect data on user behavior but computes the data it needs to perform its music personalization operations. Users listen to music and the system produces 'playcounts' computing artist names if particular conditions are met. Users of Last.fm will see artist data just after they have been filtered, organized, mapped and ordered by the operations of 'similarity network' and 'autocorrection system'. Interestingly, here users will also see each other in relation to artist visibility. For instance, a user will be able to see another user more easily if they are both listeners of a particular artist. Thanks to how the computational operations of recommender systems technology work, networks of artists and networks of users become interchangeable, flipsides of the same coin (Jannach et al., 2010; Seaver, 2012). Furthermore, when users browse content discovery features on Last.fm they just navigate the network of similar artists (and indirectly of similar users) that has been ordered by the ranking of 'similarity' score.

As our study demonstrates, artist visibility is constructed by a complex infrastructuring work of socio-technical conventions that ingest, map and clean data, at the same time establishing and reinforcing the conventions of their own data operations as the context against which the same data operations make sense (the arbitrary convention of 'similarity score' for instance as the context against which it is possible to compute similarity networks).

The importance that these abstract operations acquire for user behavior and network dynamics is reflected in our results. Indeed, 'autocorrection system' is associated with 'artist visibility' even more strongly than 'similarity network'. 'Autocorrection' is the technology of data mapping and cleansing, the infrastructural mechanism by which, as our results indicate, the more associated incorrect names, the more visible the artists. It means that the more 'autocorrection system' works by mapping incorrect artist names toward the correct one, the more visible the corresponding artists is likely to be. That is exactly what the 'autocorrection system' is supposed to do, it maps multiple entries of the same data to one entry; obviously by doing this it redirects wrong paths to a correct one. Ironically 'autocorrection' cannot correct names – database entries – on third party devices and applications. What it does is instead a mapping – linking different user databases with Last.fm database entries. By doing this, it makes artist more visible thus showing stronger association with user listening activity. The limitations of our model notwithstanding, we believe this result provides new and unique evidence that on social media platforms visibility is, first and foremost, a product of specific socio-technical infrastructural work. 'Autocorrection' is clearly a decisive infrastructural component of Last.fm, which sustains social media functioning (artist visibility emergence) and social media user behavior. 'Autocorrection' supplies a technical solution to the lack of standards of the music industry classification, constantly re-mapping music data from external and distributed databases to Last.fm system. In so doing it shapes the "visible features" of Last.fm, leads to network emergence and conditions user activities (from navigation to listening activities) and participation at the interface level.

The iterative mechanism of boundary making

The previous section shows how Last.fm data infrastructure is formed by a system of decentralized and distributed databases. Last.fm re-centralizes, filters and computes data and content produced on other platforms and devices on top of which it produces its own social

data. We have theorized the role of APIs as boundary resources as the second generative mechanism. Enabling coordination of activities and data, APIs tie together heterogeneous actors while negotiating the openness and closure of the infrastructure. Our findings support our theorizing that sees the role of APIs as central to the very functioning of social media. We also find that the importance of APIs is so central to social media that enters in their very definition.

For instance, we find that, in order to have the possibility to listening to music, Last.fm's users need to upload or to attach videos from YouTube. Our empirical results show that this movement of content across platforms and layers enabled by API has the strongest association to user listening activity. Probably unsurprisingly the result is significant. Beyond its statistical meaning, to us it speaks of current developments of social media networks and the complex – albeit invisible – role that social technologies such as APIs, and social data networks enact in todays' data economy. Last.fm is devoid of prepackaged content and it is built on top of networks of social data, their procurement and reuse (listening behavior procured from user platform participation on external social media platforms). In turn, the social data Last.fm computes after many other complex processes (the ingestion, aggregation, cleansing, and mapping) are used to suggest personally relevant music to each and every user. It is perhaps even more interesting is the fact that in our model this is the variable that appears to be most significantly connected to user listening activity. Beside Last.fm social functionalities and recommender systems advancements what actually really counts to boost user listening activity can be found only by connecting to another social media. This controlled openness of social media based largely on data and content flowing is one of the most important aspects of today's social media evolution and the main findings of our research. The role of APIs supports our view of social media as infrastructures primarily designed through the standardization of interfaces (API) and protocols (Hanseth, 2001). An

important implication of this finding is that on social media success will not be so much achieved by a centralized design decisions or controlled by a centralized management (Ciborra, 2000) but sustained by precise architectural principles that will produce new forms of value from data "controlled" production and exchange.

(Infra)structuring the participation of lay publics

Our last generative mechanism has been adapted by Zittrain's generativity. He saw generativity as the capacity of a system of producing unanticipated change by relying on the participation of broad and varied audiences (Zittrain, 2008). As indicated, a point seldom considered in the study of social media networks is the precise nature of the data flowing in the "social" network. Very little has been published on how the process of infrastructuring structures data and content through the standardization of the participation of lay publics. Indeed, there is no empirical evidence linking new forms of user participation to social technologies or emergent system functionalities. Drawing from these conceptual concerns, our model provides new evidence on the links between content and user behavior. We consider two aspects related to content: user generation of content (UGC) and user interaction (in particular commenting on artist pages, one of the social features of the platform).

The two variables associated with UGC are statistically insignificant. That is, the quantity of content produced by user on artists does not have a significant impact on weekly listening activity. It might be rightfully argued that this is because a quantitative analysis considers just the quantity of content and cannot take into consideration its main aspect that is its quality (for instance how informative or relevant the content produced is). We nonetheless view the fact that both our two UGC variables are statistically insignificant, as suggesting that the role of user generation of content on social media may be less important than what has been assumed so far. The role of UGC for Last.fm functioning indicates an ambiguity that seems to be a general characteristic of the role of UGC on social media. What is content on Last.fm?

If we follow the "content industry" taxonomy and include Last.fm in it, therefore limiting our notion of content here to music, then UGC just provides data and metadata to the prepackaged content of the music industry. What users generate is in the final analysis an enhancement of database records that our results show has low significance for user listening activity but may have strong significance for system operations.

Evidence from our third variable attests to the importance of social features on social media platforms and it also lends support to our point about UGC. User interaction under the form of commenting is statistically significant for user listening activity. The quantity of comments on artist profiles is likely to boost weekly user listening activity. To us, this is a central point that calls for putting content, the nature and technologies of content production back into the social media research agenda. The distinction between "network-oriented" and "content-oriented" social media networks is untenable (Berger et al., 2014; Sundararajan et al., 2013). It is theoretically very problematic to classify networks based on typologies of content without understanding how technologies of content production are tied to the nature of digital objects and standardization of social participation (Kallinikos et al., 2013). Is commenting content production? Our results show a discrepancy between the significance of UGC and user interaction (under the form of commenting) for user listening activity, which indicates that the formalization of social interaction and participation on social media has little to do with the classic distinction between "content" and "networking" activity. Last.fm in this respect is aligned with what seems a more general trend in social media evolution: the simplification of the routes along which user activities take place. Social media are large scale, distributed, interconnected and layered infrastructures. The more data and content need to flow across layers and applications the more social media necessitate agreed upon standards for data and content transmission (Bowker and Star, 2000). Once again Facebook

and its “Likes” are good exemplifications of this tendency: an abstract enough data token that is standardizing a set of related online user behaviors.

Conclusion and suggestions for further research

Our study unravels a complex infrastructure of social technologies that constantly tunes and is tuned by self-reinforcing cycles of data production, technologies of boundary making and user participation, which becomes standardized along formalized routes. Although our paper contributes significantly to address the gap in social media theorizing by providing an infrastructural frame, we are aware that further research is needed. Our own investigation just scratches the surface of the vertically distributed system of data flows that feeds the Last.fm database. Further research will be required to achieve a more comprehensive understanding of the complex vertical and horizontal architectures of data and devices that sustain social media functioning.

Social media are much more than networks, they are layered architectures of technologies of data ingestion, filtering, mapping and computation on top of which other layers of social technologies enable networks of social data flowing within and across platforms. On Last.fm this complex functioning is made possible by the system of APIs that regulate data exchange within and across platforms. APIs make up a very complex underlying grid of filters, which heavily conditions social data circulation and commercialization within and across social media networks. Understanding the role of these boundaries technologies (Ghazawneh and Henfridsson, 2010; Tilson et al., 2010; Yoo et al., 2010) is essential to analyze how do they partake in controlling what flows within and across platforms (Bucher, 2013; Constantinides, 2012). Our study acknowledge the importance of APIs for Last.fm functioning but more research is needed to explore the variety of use of these social technologies, the variety of possible platform configurations they may lead to and their co-evolution. Many of the most

innovative social media business ventures (for instance applications) are made possible by the rapid development of these APIs enabled socio-technical configurations (Eaton et al., 2015).

Our study contributes to the literature on social media and social media networks by highlighting what appears to us as an ongoing evolution of social media that seems to place them in-between platforms and infrastructures and that conditions the even more interesting evolution of the web from a transaction based network to a social web.

Several authors have noted the current trend toward what has been called the "platformization of the web" or the "appliancization of the web" (van Dijck, 2013; Zittrain, 2008). We have referred to this important theme throughout our paper by mentioning social media as infrastructures or the infrastructuring of social media. If we are right and social media is infrastructuring the web then the major disruption would not be at the level of software or platform but in the redefinition of roles and behaviors (Bendik Bygstad, 2010; Henfridsson and Bygstad, 2013).

Although the object of our study is not the social web, we contribute to this interesting and timely strand by pointing out to the interconnectedness of social media platforms and its transformative effects. Social media are complex layered machineries of social data production and computation that seem to assume different configurations at different levels. Benefiting from a certain model of social participation their functioning is conditioning what others, not social media websites are doing. In our case of Last.fm for instance, an assumed incompatibility of 'audio.scrobble' with a certain listening device would condition a Last.fm user to abandon that particular device (or to abandon Last.fm). Without further research and analysis at the level of data production, circulation and computation and their implications for user behavior the innovativeness and role of social media for the digital economy get lost.

Certain ontological and architectural characteristics of social media are colonizing part of the web: how users behave online, or what they expect from online services. Nowadays is impossible to sign-in to a new online service without finding a Facebook or Twitter authentication API through which users can join without filling any data fields. Platformization of the web means not only the adoption of certain technologies but also the standardization of certain related behaviors. Among these ontological traits of the making of the social web, our study highlights a certain tendency toward the simplification of user activities. Social participation is getting away from content production and it tends to rely on more abstract and exchangeable type of data production and related behavior. Further research is needed to understand to what extent social media distinctive technological features have already intermeshed with online social behavior.

Many of the findings we have shows have significant implication for business and organizations. One of the most interesting is perhaps the primary question raised by our results regarding a more general understanding of social media technological characteristics and how they define a new business model. Assuming that on our music discovery platform the content is music, essentially Last.fm is a "content-oriented site" without content. There is no content flowing in the system but just data on user listening behavior, that is, 'playback events', metadata about 'artist names' (UGC) and user comments. Oestreicher-Singer and Zamalson (2013) inadvertently arrived at similar conclusions when they show that willingness to pay on Last.fm is driven by social participation rather than content consumption. We show that things are more complex than this. On social media there is no social participation without technological mediation. Successful businesses that venture on social media need to dig into the business of data rather than in the business of social participation (if that ever exist).

Future studies can address the severe limitations of our work along several additional routes. First, the use of secondary data in our qualitative research design certainly limits the comprehensiveness and ampleness of our description and explanation of Last.fm functioning. However Last.fm and its system of APIs have offered to scholars the possibility to gather data and study its functioning to a great length. We compared our own account with other pre-existing contributions (Beer and Taylor, 2013; Oestreicher-Singer and Zalmanson, 2013) and find that our study, albeit reliant on secondary sources, shows great external validity.

Second, it is extremely relevant to point out that our quantitative research design does not want in any way to account for the causality of what we have called the generative mechanisms of social media networks. A quantitative model by its nature provides correlations only, measuring association among variables. What our model does is to complement our research design effectively, insofar as it provides rich enough evidence to prove the strong association of infrastructural elements to user behavior. Having said that, we are aware that the model has its own limitations and much more is needed to cumulate diversified empirical evidence on the association between social technology and user behavior on social media.

Third, we offer a theorization of three interrelated generative mechanisms in social media. To our knowledge, this is the first paper that analyzes a social media platform through the lens of information infrastructure. The original contribution of our research has a cost. Much more work is needed both at the empirical level and at the conceptual level. At the empirical level the challenge is to provide a richer variety of cases of social media as infrastructures. Do social media infrastructural conditions vary across social media types? Do they exhibit the same set of generative mechanisms? At the conceptual level the challenge is to refine our understanding of social media as infrastructure, but also even more so to be able to add something new to the already rich information and digital infrastructure literature. To what

extent do social media show characteristics that have not been observed in other information infrastructures? This generative potential of social media as infrastructure is, we believe, what make them so interesting to study.

References

Aaltonen, A. and N. Tempini (2014). "Everything Counts in Large Amounts: A Critical Realist Case Study in Data-based Production" *Journal of Information technology* 29 (1), 97-110.

Adomavicius, G. and A. Tuzhilin (2005) "Toward the next generation of the recommender systems: A survey of the state-of-the-art and possible extension" *IEEE Transactions on Knowledge and Data Engineering* 17 (6), 734-749

Agarwal, R. and V. Dhar (2014) "Big data, data science, and analytics: The opportunity and challenge for IS research" *Information Systems Research* 25 (3), 443-448

Alaimo, C. and J. Kallinikos (2016) "Encoding the everyday: The infrastructural apparatus of social data" In *Big Data Is Not a Monolith*, MIT Press, ed Sugimoto, C.R., Ekbia, H.R. and M. Mattioli

Allan, J., Carbonell, J., Doddington, G., Yamron, J., Y. Yang (1998) "Topic detection and tracking pilot study final report." In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*

Anand, S.S. and B. Mobasher (2005) "Intelligent techniques for web personalization." *Lecture Notes in Computer Science* 3169, 1-36

Anderson, C. (2006). *The Long Tail: Why the Future of Business is Selling Less of More.* 1st Edition. New York: Hyperion.

Andrejevic, M. (2011) "Surveillance and Alienation in the Online Economy" *Surveillance & Society* 8 (3), 278-287

Aral, S. and D. Walker (2011). "Creating social contagion through viral product design: A randomized trial of peer influence in networks." *Management science*, 57(9), 1623-1639.

Aral, S. and D. Walker (2012) "Identifying influential and susceptible members of social networks." *Science* 337 (6092), 337-341

Aral, S., Dellarocas, C. and D. Godes (2013). "Introduction to the special issue-social media and business transformation: A framework for research." *Information Systems Research* 24(1), 3-13.

Aral, S., Muchnik, L., and A. Sundararajan (2009) "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks" *PNAS* 106 (51), 21544-21549

Aramaki, E., Maskawa, S., and M. Morita (2011) "Twitter catches the flu: detecting influenza epidemics using Twitter." In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*

Archer, M. S. and R. Bhaskar (1998). *Critical realism: essential readings.* Routledge

Arvidsson, A. (2006) "'Quality singles': internet dating and the work of fantasy" *New Media & Society* 8 (4), 671-690

Asur, S., Huberman, B, et al. (2010) "Predicting the future with social media" In *Proceedings of International Conference on Web Intelligence and Intelligent Agent Technology*

Axelrod, R. (1997) *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration.* Princeton, New Jersey: Princeton University Press.

Baeza-Yates, R. and B. Ribeiro-Neto (2011) *Modern Information Retrieval.* New York: ACM Press

Baker D.A. and G.P. Algorta (2016) "The Relationship Between Online Social Networking and Depression: A Systematic Review of Quantitative Studies." *Cyberpsychology, Behavior and Social Networking* 19 (11), 638-648

Balabanovic, M. and Y. Shoham (1997). "Fab: content-based collaborative recommendation." *Communications of the ACM* 40 (3), 66-72

Bapna, R., and A. Umyarov (2012). "Are Paid Subscriptions on Music Social Networks Contagious? A Randomized Field Experiment." *SOBACO Working Paper.* Carlson School of Management, University of Minnesota.Gjoka et al. (2010)

Barabasi, A. L. (2002). *Linked: The New Science of Networks.* Perseus Pub.

Bateman, P. J., Gray, P. H. and B. S. Butler (2011). "The Impact of Community Commitment on Participation in Online Communities." *Information Systems Research* 22 (4), 841-854.

Baym, N. (2015) "Social Media and the Struggle for Society" *Social Media + Society* April-June, 1-2

Beer, D. (2009) "Power through the algorithm? Participatory web cultures and the technological unconscious" *New Media & Society* 11 (6), 985-1002

Beer, D. and M. Taylor (2013) "The Hidden Dimensions of the Musical Field and the Potential of the New Social Data." *Sociological Research Online* 18(2), 14.

Belkin, N.J. and W.B. Croft (1992) "Information Filtering and Information Retrieval: Two Sides of the Same Coin?" *Communication of the ACM* 35 (12), 29-38

Bello-Orgaz, G., Jung, J.J., D. Camacho (2016) "Social big data: Recent achievements and new challenges" *Information Fusion* 28, 45-59

Ben, L. (2007) "Introducing Masculinity Studies to Information Systems research: The case of Gaydar" *European Journal of Information Systems* 16 (5), 658-665

Benkler, Y. (2006) *The Wealth of Networks. How Social Production Transforms Markets and Freedom.* New Haven, CT: Yale University Press

Bennett, C.J. (2015) "Trends in voter surveillance in western societies: Privacy intrusion and democratic implications" *Surveillance & Society* 13 (3/4), 370-384

Berger, K., Klier, J., Klier, M. and F. Probst (2014) "A Review of Information Systems Research on Online Social Networks." *Communications of the Association for Information Systems* 35(1), 8.

Berger, K., Klier, J., Klier, M. and F. Probst (2014) "A Review of Information Systems Research on Online Social Networks." *Communications of the Association for Information Systems* 35(1), 8.

Bettman, J.R. and M.A. Zins (1979). "Information format and choice task effects in decision making." *Journal of Consumer Research* 6 (2), 141-153.

Bettman, J.R. and P. Kakkar (1977). "Effects of information presentation format on consumer information acquisition strategies." *Journal of Consumer Research* 3 (4), 233-240.

Birkbak, A. and H.B. Carisen (2016) "The world of Edgerank: Rhetorical justifications of Facebook's News Feed algorithm" *Computational Culture* 5

Bivens and Haimson (2016) "Baking Gender into Social Media Design: How Platforms Shape Categories for Users and Advertisers" *Social Media + Society* October-December, 1-12

Boerman, S.C. and S. Kruikeimeter (2016) "Consumer responses to promoted tweets sent by brands and political parties" *Computers in Human Behavior* 65, 285-294

Borgatti, S. P., Mehra, A., Brass, D. J. and G. Labianca (2009) "Network analysis in the social sciences." *Science* 323(5916) 892-895.

Bowker, G. C. and S.L. Star (2000) *Sorting things out: classification and its consequences.* MIT Press.

boyd, d. and N.B. Ellison (2008) "Social Network Sites: Definition, History, and Scholarship." *Journal of Computer-Mediated Communication* 13, 210-230.

Braun, J. (2015) "Social Media and Distribution Studies" *Social Media + Society* April-June, 1-2

Brookes, T. (2014a). "Descriptive Metadata in the Music Industry: Why It Is Broken and How to Fix it - Part One." *Journal of Digital Media Management* 2 (3), 263-282.

Brookes, T. (2014b). "Descriptive Metadata in the Music Industry: Why It Is Broken and How to Fix it - Part Two." *Journal of Digital Media Management* 2 (4), 359-374.

Brynjolfsson, E. and A. McAfee (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* W. W. Norton.

Brynjolfsson, E., Kim, S. T. and J. Oh (2013). "User Investment and Firm Value: Case of Internet Firms." In: *Workshop for Information Systems and Economics (WISE) 2013*

Buchanan, T. (2015) "Aggressive priming online: Facebook adverts can prime aggressive cognitions" *Computers in Human Behavior* 48, 323-330

Bucher, T. (2012) "Want to be on the top? Algorithmic power and the threat of invisibility on Facebook" *New Media & Society* 14 (7), 1164-1180

Bucher, T. (2013) "Objects of intense feeling: The case of the Twitter API" *Computational Culture*

Bucher, T. (2015) "Networking, or What the Social Means in Social Media" *Social Media + Society*, April-June, 1-2

Burt, R.S. (1992) *Structural Holes: The Social Structure of Competition.* Cambridge, MA: Harvard University Press

Butler, B.S., Bateman, P.J., Gray P.H. and E.I. Diamant (2014) "An attraction-selection-attrition theory of online community size and resilience" *MIS Quarterly* 38 (3), 699-728

Bygstad, B., Munkvold, B. E. and O. Volkoff (2016) "Identifying generative mechanisms through affordances: a framework for critical realist data analysis." *Journal of Information Technology* 31, 83-96.

Caers, R., Feyter, T.D., et al. (2013) "Facebook: A Literature review" *New Media & Society* 15 (6), 982-1002

Carah, N. and A. Dobson (2016) "Algorithmic hotness: Young women's 'promotion' and 'reconnaissance' work via social media body images" *Social Media + Society* October-December, 1-10

Cardullo, P. (2015) "'Hacking multitude' and Big Data: Some Insights from the Turkish 'digital coup'" *Big Data & Society* January-June, 1-14

Carr, C.T. Wohn, D.Y. and R.A. Hayes (2016b) "It's the audience: Difference in social support across social media" *Social Media + Society* October-December, 1-12

Carr, C.T., Wohn, D.Y. and R.A. Hayes (2016a) "Relational closeness, automaticity and interpreting social support from paralinguistic digital affordances in social media" *Computers in Human Behavior* 62, 385-393

Celma, O. (2008). *Music Recommendation and Discovery in the Long Tail.* PhD thesis. Universitat Pompeu Fabra.

Celma, O. and P. Cano (2008) "From hits to niches? Or how popular artists can bias music recommendation and discovery." In *2nd Netflix-KDD Workshop*

Celma, O. and P. Lamere (2011). "If You Link Radiohead, You Might Like This Article." *AI Magazine* 32 (3), 57-66

Chainey, S., Tompson, L., and S. Uhlig (2008) "The utility of hotspot mapping for predicting spatial patterns of crime" *Security Journal* 21 (1), 4-28

Chakrabarti, S. (2002) *Mining the Web: Discovering knowledge from hypertext data.*

Chen, J., Xu H. and A.B. Winston (2011) "Moderated online communities and quality of user-generated content" *Journal of Management Information Systems* 28 (2), 237-268

Chen, Y., Boring, S. and A. Butz (2010). "How Last.fm Illustrates the Musical World: User Behavior and Relevant User-Generated Content." In *International Workshop on Visual Interfaces to the Social and Semantic Web (VISSW) 2010*

Chmiel, A., Sobkowicz, P., Sienkiewicz, J., Paltoglou, G., Buckley, K., Thelwall, M., and J. A. Holyst (2011). "Negative Emotions Boost Users Activity at BBC Forum." *Physica A: Statistical Mechanics and its Application* 390 (16), 2936-2944.

Ciborra, C. (2000) *From Control to Drift: The Dynamics of Corporate Information Infastructures.* Oxford University Press.

Claussen, J., Kretschmer, T. and P. Mayrhofer (2013) "The effects of rewarding user engagement - The case of Facebook apps" *Information Systems Research* 24 (1), 186-200

Colace, F., de Santo, M.D., Greco, L., Moscato V. and A. Picariello (2015) "A collaborative user-centered framework for recommending items in online social networks." *Computers in Human Behavior* 51, 694-704

Constantiou, I. and J. Kallinikos (2015a) "New games, new rules: big data and the changing context of strategy" *Journal of Information Technology* 30 (1), 44-57

Constantiou, I. and J. Kallinikos (2015b) "Big data revisited: A rejoinder" *Journal of Information Technology* 30 (1), 70-74

Conte, R., Gilbert, N., et al. (2012) "Manifesto of computational social science." *European Physical Journal* 214, 325-346

Couldry, N. and J. van Dijck (2015) "Researching Social Media as if the Social Mattered" *Social Media + Society* July-December, 1-7

Crawford, K. and T. Gillespie (2016) "What is a flag for? Social media reporting tools and the vocabulary of complaint" *New Media & Society* 18 (3), 410-428

Croft, W.B. and D.J. Harper (1979) "Using Probabilistic Models of Document Retrieval without Relevance Information". *Journal of Documentation*. 35 (4), 285-295

Cross, R. and J. Cummings (2004) "Tie and network correlates of performance in knowledge intensive work." *Academy Management Journal* 47 (6), 928-937

Culotta, A. (2010) "Towards detecting Influenza epidemics by analyzing Twitter messages" In *1st Workshop on Social Media Analytics*

David, G. and C. Cambre (2016) "Screened Intimacies: Tinder and the Swipe Logic" *Social Media + Society* April-June, 1-11

Denning, P.J. (1982) "Electronic Junk". *Communications of the ACM* 25 (3), 163-165

Derntl, M., Hampel, T., Moschnig-Pitrik, R. and T. Pitner (2011) "Inclusive social tagging and its support in Web 2.0 services" *Computers in Human Behavior* 27, 1460-1466

Dou, W., Lim, K.H., Su, C., Zhou N. and N. Cui (2010). "Brand Positioning Strategy Using Search Engine Marketing." *MIS Quarterly* 34 (2), 261-279.

Easley, D. and J. Kleinberg (2010) *Networks, Crowds, and Markets: Reasoning About a Highly Connected World.* Cambridge University Press.

Eaton, B., Elaluf-Calderwood, S., Sorensen, C. and Y. Yoo (2015) "Distributed tuning of boundary resources: the case of Apple's iOS service system." *MIS Quarterly* 39(1), 217-243.

Edwards, A., Housley, W., et al. (2013) "Digital social research, social media and the sociological imagination: surrogacy, augmentation and re-oreintation" *International Journal of Social Research Methodology* 16 (3), 245-260

Ekbia, H.R. (2009) "Digital Artifacts as Quasi-Objects: Qualification, Mediation, and Materiality" *Journal of the American Society for Information Science and Technology* 60 (12), 2554-2566

Ekstrand, M.D., Riedl, J.T. and J.A. Konstan (2010). "Collaborative Filtering Recommender Systems." *Foundations and Trends in Human-Computer Interaction* 4 (2), 81-173.

Elberse, A. (2008) "Should you invest in the long tail?" *Harvard Business Review*

Ellerbrok, A. (2010) "Empowerment: Analysing Technologies of Multiple Variable Visibility" *Surveillance & Society* 8 (2), 200-220

Ellison, N.B. and D.M. boyd (2013) "Sociality Through Social Network Sites." In *The Oxford Handbook of Internet Studies*, Oxford, UK: Oxford University Press, ed. W.H. Dutton

Elmer, G. (2012) "Live research: Twittering an election debate" *New Media & Society* 15 (1), 18-30

Elmer, G., Langlois, G. and J. Redden (2015) *Compromised Data: From Social Media to Big Data.* Bloomsbury Academic.

Emirbayer, M. and J. Goodwin (1994) "Network analysis, culture, and the problem of agency." *American Journal of Sociology* 99(6), 1411-1454.

Erkstrand, M.D., Riedl, J.T., and J.A. Konstan (2010) "Collaborative Filtering Recommender Systems"

Espeland, W.N. and M. Sauder (2007). "Ranking and Reactivity: How Public Measures Recreate Social Worlds." *American Journal of Sociology* 113 (1), 1-40

Fama, E. (1970) "Efficient capital markets: A review of theory and empirical work" *Journal of Finance* 51 (1), 55-84

Faulkner, P. and J. Runde (2011) "The Social, the Material, and the Ontology of Non-Material Technological Objects." In *27th European Group for Organizational Studies Colloquium*

Felt, M. (2016) "Social media and the social sciences: How researchers employ Big Data analytics" *Big Data & Society* January-June, 1-15

Fulk, J. and Y.C. Yuan (2013) "Location, motivation and social capitalization via enterprise social networking" *Journal of Computer-Mediated Communication* 19 (1), 20-37

Fullerton, G. (2003). "When Does Commitment Lead to Loyalty?" *Journal of Service Research* 5 (4), 333-344.

Fullerton, G. (2005). "The Service Quality - Loyalty Relationship in Retail Services: Does Commitment Matter?" *Journal of Retailing and Consumer Services* 12 (2), 99-111.

Gallagher, S.E. and T. Savage (2013) "Cross-cultural analysis in online community research: A literature review." *Computers in Human Behavior* 29, 1028-1038

Gehl, R. W. (2014). *Reverse Engineering Social Media: Software, Culture, and Political Economy in New Media Capitalism.* Temple University Press.

Gehl, R.W. (2011) "The archive and the processor: The internal logic of Web 2.0" *New Media & Society* 13 (8), 1228-1244

Gehl, R.W. (2015) "The case for alternative social media" *Social Media + Society* July-December, 1-12

Gerber, M.S. (2014) "Predicting crime using twitter and Kernel density estimation" *Decision Support System* 61, 115-125

Gerlitz, C. and A. Helmond (2013) "The like economy: Social buttons and the data-intensive web" *New Media & Society* 15 (8), 1348-1365

Ghazawneh, A. and O. Henfridsson, (2010). "Governing third-party development through platform boundary resources." In *International Conference on Information Systems (ICIS)*.

Ghose, A., Goldfarb, A. and S.P. Han (2012) "How is the mobile internet different? Search costs and local activities" *Information Systems Research* 24 (3), 1-19

Gillespie, T. (2010) "The politics of 'platforms'." *New Media & Society* 12(3), 347-364.

Gillespie, T. (2015) "Platform Intervene" *Social Media + Society* April-June, 1-2

Ginsberg, J., Mohebbi, M.H., et al. (2009) "Detecting influenza epidemics using search engine query data" *Nature* 457, 1012-1015

Goggin, G. (2014) "Facebook's mobile career" *New Media & Society* 16 (7), 1068-1086

Goldberg, K., Nochols D., Oki B.M., and D. Terry (1992) "Using Collaborative Filtering to Weave an Information Tapestry" *Communications of the ACM* 35 (12), 61-70

Goldenberg, J., Oestreicher-Singer, G. and S. Reichman (2012) "The Quest of Content: How User Generated Links Can Facilitate Online Exploration." *Journal of Marketing Research* 49 (4), 452-468

Goodwin, I., Griffin, C., et al. (2016) "Precarious popularity: Facebook drinking photos, the attention economy, and the regime of the branded self" *Social Media + Society* January-March, 1-13

Granovetter, M. S. (1973) "The strength of weak ties." *American Journal of Sociology* 78(6), 1360-1380.

Greengrass, E. (2000) "Information Retrieval: A Survey."

Greenwood, B.N. and A. Gopal (2015) "Tigerblood: Newspaper, Blogs, and the Founding of Information Technology Firms" *Information Systems Research* 26 (4), 812-828

Grosser, B. (2014) "What do metrics want? How quantification prescribes social interaction on Facebook." *Computational Culture*

Haciyakupogiu, G. and W. Zhang (2015) "Social Media and Trust during the Gezi Protests in Turkey" *Journal of Computer-Mediated Communication* 20 (4), 450-466

Hanani, U., Shapira, B. and P. Shoval (2000) "Information Filtering: Overview of Issues, Research and Systems." *User Modeling and User-Adapted Interaction*. 11 (3), 203-259

Hanseth, O. (2002) "From systems and tools to networks and infrastructures-from design to cultivation. Towards a theory of ICT solutions and its design methodology implications." Unpublished manuscript.

Hanseth, O. and K. Lyytinen (2010) "Design theory for dynamic complexity in information infrastructures: the case of building internet." *Journal of Information Technology* 25(1), 1-19.

Haythornthwaite, C. (2002) "Strong, Weak, and Latent Ties and the Impact of New Media" *The Information Society* 18, 385-401

Helmond, A. (2013) "The Algorithmization of the Hyperlink" *Computational Culture*

Helmond, A. (2015) "The Platformization of the Web: Making Web Data Platform Ready" *Social Media + Society* July-December, 2015

Henfridsson, O. and B. Bygstad (2013). "The generative mechanisms of digital infrastructure evolution." *MIS Quarterly* 37(3), 907-931.

Herlocker, J.L., Konstan, J.A., et al. (1999) "An Algorithmic Framework for Performing Collaborative Filtering". In *Proceedings of the 22nd Annual International ACM SIGIR Conference*

Heyman, R. and J. Pierson (2015) "Social Media, Delinguistification and Colonization of Lifeworld: Changing Faces of Facebook." *Social Media + Society* July-December, 1-11

Hill, W., Stead, L., Rosenstein, M., and G. Furnas (1995) "Recommending and evaluating choices in a virtual community use". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*

Hochman, N. (2014) "The social media image" *Big Data & Society* July-December, 1-15

Hogan, N, (2015) "Data flows and water woes: The Utah Data Center" *Big Data & Society* July-December, 1-12

Hosanagar, K., Fleder, D., Lee D. and A. Buja (2013). "Will the Global Village Fracture into Tribes? Recommender Systems and their Effects on Consumer Fragmentation." *Management Science* 60 (4), 805-823.

Huber, O. (1980). "The influence of some task variables on cognitive operations in an information processing decision model." *Acta Psychologica* 45 (1-3), 187-196.

Ilten, C. (2015) "'Use Your Skills to Solve This Challenge!': The Platform Affordances and Politics of Digital Microvolunteering" *Social Media + Society* July-December, 1-11

Introna, L. D. and H. Nissembaum (2000) "Shaping the Web: Why the Politics of Search Engines Matters." *The Information Society* 16(3), 169-185.

Jannach, D., Zanker, M., Felfernig, A. and G. Friedrich (2011) *Recommender Systems: A Introduction.* Cambridge, UK: Cambridge University Press

Jansen, B.J., Zhang, M., Sobel, K. and A. Chowdury (2009) "Twitter power: tweets as electronic word of mouth" *Journal of the Association for Information Science and Technology* 60 (11), 2169-2188

Jarvenpaa, S.L. (1989). "The effect of task demands and graphical format on information processing strategies." *Management Science* 35 (3), 175-189.

Jennings, D. (2007) *Net, Blogs and Rock 'n' Roll: How Digital Discovery and What it Means for Consumers.* 1st Edition, London: Nicholas Brealey Publishing.

Kaldrack, I. and T. Rohle (2014) "Divide and Share: Taxonomies Orders and Masses in Facebook's Open Graph" *Computational Culture*

Kallinikos, J. and J,-C. Mariategui (2011) "Video as Digital Object: Production and Distribution of Video Content in the Internet Media Ecosystem" *The Information Society* 27 (5), 281-294

Kallinikos, J. and N. Tempini (2014) "Patient data as medical facts: social media practices as a foundation for medical knowledge creation" *Information Systems Research* 25 (4)

Kallinikos, J., Aaltonen, A and A. Marton (2010) "A Theory of Digital Objects" *First Monday* 15 (6)

Kallinikos, J., Aaltonen, A. and A. Marton (2013) "The Ambivalent Ontology of Digital Artifacts." *MIS Quarterly* 37(2), 357-370.

Kane, G. C., Alavi, M., Labianca, G. and S.P. Borgatti (2014). "What's different about social media networks? a framework and research agenda." *MIS Quarterly* 38(1), 275-304.

Kaplan, A.M., and M. Haenlein (2010). "Users of the world, unite! The challenges and opportunities of Social Media." *Business horizon* 53 (1), 59-68

Kennedy, H. and G. Moss (2015) "Known or knowing publics? Social media data mining and the question of public agency" *Big Data & Society* July-December, 1-11

Khosravi, P., Rezvani, A. and A. Weiwiora (2016) "The impact of technology on older adults' social isolation" *Computers in Human Behavior* 63, 594-603

Kim, D.H., Sung, Y.H., et al. (2016) "Are you on Timeline or News Feed? The roles of Facebook pages and construal level of increasing ad effectiveness" *Computers in Human Behavior* 57, 312-320

Kim, H-S and A. Mrotek (2016) "A functional and structural diagnosis of online health communities sustainability: A focus on resource richness and site design features" *Computers in Human Behavior* 63, 362-372

Kleinmuntz, D.N. and D.A. Schkade (1993). "Information Displays and Decision Processes.", *Psychological Science* 4 (4), 221-227

Knorr-Cetina, K. (2001) "Objectual practice" In *The practice turn in contemporary theory*, Abingdon: Routledge, ed Shatzki, T.R., Von Savigny, E., and K. Knorr-Cetina.

Knox, H., Savage, M. and P. Harvey (2006). "Social networks and the study of relations: networks as method, metaphor and form." *Economy and Society* 35(1), 113-140.

Konstan, J. A., and J. Riedl (2012). *Deconstructing Recommender Systems*. URL: http://spectrum.ieee.org/computing/software/deconstructing-recommender-systems (visited on 11/21/2014).

Krijnen, D., Bot, R. and G. Lampropoulos (2014) "Automated Web Scraping APIs"

Kruikemeier, S. Sezgin, M. and S.C. Boerman (2016) "Political Microtargeting: Relationship Between Personalized Advertising on Facebook and Voters' Responses" *Cyberpsychology, Behavior, and Social Networking* 19 (6), 367-372

Ku, C.-H., and G. Leroy (2014) "A decision support system: automated crime report analysis and classification for e-government" *Government Information Quarterly* 31 (4), 534-544

Kumpel, A.S., Karnowski, V. and T. Keyling (2015) "News Sharing in Social Media: A Review of Current Research on News Sharing Users, Content, and Networks" *Social Media + Society* July-December, 1-14

Kwoka, J. E. (1985). "The Herfindahl Index in Theory and Practice." *The Antitrust Bulletin* 30, 915-947

Lopez, J. and J. Scott (2000) *Social structure.* Open University Press.

Langlois, G. (2015) "What are the stakes in doing critical research on social media platforms?" *Social Media + Society* April-June, 1-2

Lazer, D., Kennedy, R., King, G., and A. Vespignani (2014) "The parable of Google Flu, traps in big data analysis" *Science* 343 (6176), 1203-1205

Lazer, D., Pentland, A.S., et al. (2009) "Life in the network: the coming age of computational social science" *Science* 323(5915) 721-723

Lee-Won, R.J., Abo, M.M., Na, K. and T.N. White (2016) "More than numbers: Effects of social media virality metrics on intention to help unknown others in the context of bone marrow donation" *Cyberpsychology, Behavior and Social Networking* 19 (6), 404-411

Leonardi, P.M. (2014) "Social media, knowledge sharing, and innovation: Toward a theory of communication visibility" *Information Systems Research* 25 (4), 796-816

Leonardi, P.M. (2015) "Ambient awareness and knowledge acquisition: Using social media to learn 'who knows what' and 'who knows whom'" *MIS Quarterly* 39 (4), 747-762

Leonardi, P.M., Huysman, M. and C. Steinfield (2013) "Enterprise Social Media: Definition, History, and Prospects for the Study of Social Technologies in Organizations." *Journal of Computer-Mediated Communication* 19 (1), 1-19

Lessig, L. (2006) *Code: Version 2.0*. New York: Basic Books

Levina, N. and M. Arriaga (2014) "Distinction and status production on user-generated content platforms: Using Bourdieu's theory of cultural production to understand social dynamics in online fields." *Information Systems Research* 25(3), 468-488.

Levy, M. and K. Bosteels (2010). "Music Recommendation and the Long Tail." In *Workshop on Music Recommendation and Discovery (Womrad) 2010*.

Li, L., Gao, P. and J. Mao (2014) "Research on IT in China: a call for greater contextualization." *Journal of Information Technology* 29 (3), 208-222

Lin, N. (2001) *Social Capital: A Theory of Social Structure and Action.* Cambridge, UK: Cambridge University Press

Linden, G., Smith, B. and J. York (2003) "Amazon.com recommendations: item-to-item collaborative filtering". *Internet Computing, IEEE* 7 (1), 76-80

Ling, K., Beenen, G., et al. (2005) "Using social psychology to motivate contributions to online communities" *Journal of Computer-Mediated Communication* 10 (4)

Lovink, G. and M. Rasch (2013). *Unlike us reader: social media monopolies and their alternative.* Amsterdam: Institute of Network Cultures.

Luhn, H.P. (1958) "A Business Intelligence System". *IBM Journal of Research and Development* 2 (4), 314-319

Lukyanenko, R., Parsons, J. and Y. Wiersma (2014) "The IQ of the Crowd: Understanding and Improving Information Quality in Structured User-Generated Content" *Information Systems Research* 25 (4), 669-689

Luo, X., Zhang, J. and W. Duan (2013) "Social Media and Firm Equity Value" *Information Systems Research* 24 (1), 146-163

Lynn, W. (2013) "Social Network Effects on Productivity and Job Security: Evidence from the Adoption of a Social Networking Tool" *Information Systems Research* 24 (1), 30-51

Ma, M. and R. Agarwal (2007) "Through and Glass Darkly: Information Technology Design, Identity Verification and Knowledge Contribution in Online Communities" *Information Systems Research* 18 (1), 42-67

MacKee, F. (2016) "Social media in gay London: Tinder as an alternative to hook-up apps" *Social Media + Society* July-September, 1-10

Majchrzak, A., Faraj, S., Kane, G.C. and B. Azad (2013) "The contradictory influence of social media affordances on online communal knowledge sharing" *Journal of Computer-Mediated Communication* 19 (1), 38-55

Malinen, S. (2015) "Understanding user participation in online communities: A systematic literature review of empirical studies" *Computers in Human Behavior* 46, 228-238

Malone, T.W., Grant, K.R., Furbak, F.A., Brobst S.A. and M.D. Cohen (1987) "Intelligent Information Sharing Systems" *Communications of the ACM* 30(5), 390-402

Manovich, L. (2001) *The Language of New Media.* Cambridge MA: MIT Press.

Manovich, L. (2011) "Trending: The promises and the challenges of big social data." In *Debates in the Digital Humanities*, University of Minnesota Press, ed. M.K. Gold

Marchionini, G. (1995) *Information Seeking in Electronic Environments*. Cambridge University Press

Marres, N. and E. Weltevrede (2013) "Scraping the social? Issues in live social research." *Journal of Cultural Economy* 6 (3), 313-335

McPhee, W.N. (1963) *Formal theories of mass behavior*. Free Press

McPherson, M., Smith-Lovin, L. and J.M. Cook (2001) "Birds of a feather: Homophily in social networks." *Annual review of sociology* 27, 415-444.

Milan, S. (2015) "When algorithms shape collective action: Social media and the dynamics of cloud protesting" *Social Media + Society* July-December, 1-10

Miles, M. B. and A.M. Huberman (1994). *Qualitative data analysis: an expanded sourcebook. Sage Publications.*

Mingers, J. (2001) "Combining IS research methods: towards a pluralist methodology." *Information Systems Research* 12(3), 240-259.

Mitchell, T.M. (1997) *Machine Learning* New York: McGraw-Hill

Moreno, M.A., D'Angelo, J. and J. Whitehill (2016) "Social Media and Alcohol: Summary of Research, Intervention Ideas and Future Study Directions." *Media and Communication* 4 (3), 50-59

Morris, J. W. (2012). "Making Music Behave: Metadata and the Digital Music Commodity." *New Media & Society* 14 (5), 850-866.

Morville, P. and L. Rosenfeld (2006). *Information Architecture for the World Wide Web: Designing Large-Scale Web Sites.* 3rd Edition, California: O'Reilly Media.

Oakley, A. (2016) "Disturbing Hegemonic Discourse: Nonbinary Gender and Sexual Orientation Labeling on Tumblr" *Social Media + Society* July-September, 2016

Oard, D.W. (1997) "The State of the Art in Text Filtering." *User Modeling and User-Adapted Interaction.* 7 (3), 141-178

Oestreicher-Singer, G. and L. Zalmanson (2013) "Content or Community? A Digital Business Strategy for Content Providers in the Social Age" *MIS Quarterly* 37 (2), 591-616

Oestreicher-Singer, G. and A. Sundararajan (2012a) "Recommendation Networks and the Long Tail of Electronic Commerce." *MIS Quarterly* 36 (1), 65-83

Oestreicher-Singer, G. and A. Sundararajan (2012b). "The Visible Hand? Demand Effects of Recommendation Networks in Electronic Markets." *Management Science* 58 (11), 1963-1981.

Oh, O., Eom, C. and H.R. Rao (2015) "Role of social media in social change: An analysis of collective sense making during the 2011 Egypt revolution" *Information Systems Research* 26 (1), 210-223

Oinas-Kukkonen, H., Lyytinen, K. and Y. Yoo (2010) "Social Networks and Information Systems: Ongoing and Future Research Streams." *Journal of Association of Information Systems* 11, 61-68

Page, L., Brin, S., Motwani, R., and T. Winograd (1999) "The PageRank citation ranking: Bringing order to the web". Technical Report. Stanford InfoLab

Papacharissi, Z. (2009) "The virtual geographies of social networks: a comparative analysis of Facebook, LinkedIn and ASmallWorld" *New Media & Society* 11(1&2), 199-220

Paquet, C., Coulombier, D., Kaiser, R., and M. Ciotti (2005) "Epidemic intelligence: a new framework for strengthening disease surveillance in Europe" *European Surveillance* 11 (12), 212-214

Pazzani, M.J. and Billsus, D. (2007) "Content-based recommendation systems" In *The Adaptive Web*, Springer, ed. Brusilovsky, P., Kobsa, A., and W. Nejdl

Pietrobruno, S. (2013) "YouTube and the social archiving of intangible heritage" *New Media & Society* 15 (8), 1259-1276

Pollock, T.G. and V.P. Rindova (2003) "Media legitimation effects in the market for initial public offerings." *Academy Management Journal* 46 (5), 631-642

Poor, N. (2005) "Mechanisms of an Online Public Sphere: The Website Slashdot" *Journal of Computer-Mediated Communication* 10 (2)

Porter, A.J. and I. Hellstein (2014) "Investigating participatory dynamics through social media using a multideterminant 'frame' approach: the case of Climategate on YouTube" *Journal of Computer-Mediated Communication* 19 (4), 1024-1041

Porter, M. (1980) "An Algorithm for Suffix Stripping." *Program* 14 (3), 130-137

Power, M. (1994). *The Audit Explosion*. London: Demos

Preis, T. and H.S. Moat (2014) "Adaptive nowcasting of influenza outbreaks using Google searches" *Royal Society Open Science*

Probst, F., Grosswiele, D.-K. L. and D.-K.R. Pfleger (2013) "Who will lead and who will follow: Identifying Influential Users in Online Social Networks." *Business & Information Systems Engineering* 5(3), 179-193.

Rains, S.A. and S.R. Brunner (2014) "What can we learn about social network sites by studying Facebook? A call and recommendations for research on social network sites" *New Media & Society* 17, 114-131.

Ren, Y., Harper, F.M., et al. (2012) "Building member attachment in online communities: Applying theories of group identity and interpersonal bonds" *MIS Quarterly* 36-3, 841-864

Resnick, Iacovou, N., Suchak, M., Bergstorm P. and J. Riedl (1994) "Grouplens: An open architecture for collaborative filtering netnews." In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*

Rhoades, S. A. (1993). "The Herfindahl-Hirschman Index." *Federal Reserve Bulletin* 79 (3), 188-189.

Ridder, S.D. (2015) "Are digital media institutions shaping youth's intimate stories? Strategies and tactics in the social networking site Netlog" *New Media & Society* 17 (3), 356-374

Riedl, J. and B. Smyth (2011) "Introduction to special issue on recommender systems." *ACM Transactions on the Web* 5(1), 1-2.

Riordan, C. and H. Sorensen (1997) "Information Filtering and Retrieval: An Overview"

Rose, J. and S. Oystein (2010) "Designing Deliberation Systems" *The Information Society* 26 (3), 228-240

Salender, L. and S.L. Jarvenpaa (2016) "Digital action repertoires and transforming a social movement organization" *MIS Quarterly* 40 (2), 331-352

Salton, G. and C. Buckley (1988) "Term-weighting Approaches in Automatic Text Retrieval." *Information Processing & Management* 24 (5), 513-523

Salton, G., Wong, A., and C.S. Yang (1975) "A vector space model for information retrieval" *Journal of the American Society of Information Science* 18 (11), 613-620

Samuelson, P.A. (1965) "Proof that properly anticipated prices fluctuate randomly" *Industrial Management Review* 6 (2), 41-49

Sandvig, C. (2015) "The Social Industry" *Social Media + Society* April-June, 1-4

Sarwar, B., Karypis, G., Konstan, J., and J. Riedl (2001) "Item-based collaborative filtering recommendation algorithms" In *Proceedings of the 10th International Conference of the World Wide Web*

Savage, M. and R. Burrows (2007) "The Coming Crisis of Empirical Sociology" *Sociology* 41 (5), 885-899

Savage, M. and R. Burrows (2009) "Some further reflections on the coming crisis of empirical sociology" *Sociology* 43 (4), 762-772

Sayer, A. (1992) *Method in social science: A realist approach.* London: Routledge

Schkade, D.A. and D.N. Kleinmuntz (1994). "Information Displays and Choice Processes: Differential Effects of Organization, Form and Sequence." *Organizational Behavior and Human Decision Processes* 57 (3), 319-337

Schnapp, J. and P. Presner (2009) "Digital Humanities Manifesto 2.0"

Shah, N. (2015) "When machines speak to each other: unpacking the 'social' in 'social media'" *Social Media + Society* April-June, 1-3

Shapiro, C. and H.R. Varian (1999) *Information Rules: A Strategic Guide to the Network Economy.* Cambridge MA: Harvard Business Press

Shardanand, U. and P. Maes (1995) "Social information filtering: Algorithms for automating 'word of mouth'". In *Proceedings of the ACM Conference on Human Factors in Computing Systems*

Shaw, R. (2015) "Big Data and Reality" *Big Data & Society* July-December, 1-4

Shi, Z., Rui, H. and A.B. Whinston (2013) "Content sharing in a social broadcasting environment: evidence from twitter." *MIS Quarterly* 38(1), 123-142.

Shirky, C. (2008). *Here Comes Everybody: The Power of Organizing Without Organizations.* 1st Edition. New York: Penguin Books.

Simon, H.A. (1969) *The Sciences of the Artificial*. Cambridge, MA: MIT Press

Skageby, J. (2009) "Exploring Qualitative Sharing Practices of Social Metadata: Expanding the Attention Economy" *The Information Society* 25, 60-72

Skageby, J. (2010) "Gift-giving as a conceptual framework: framing social behavior in online networks." *Journal of Information Technology* 25, 170-177

Smith, B. (1996) *On the origin of objects.* Boston: MIT Press.

Spagnoletti, P., Resco A. and G. Lee (2015) "A design theory of digital platforms supporting online communities: a multiple case study" *Journal of Information Technology* 30, 364-380

Spears, R. and M. Lea (1992) "Social influence and the influence of the 'social' in computer-mediated communication" In *Contexts of computer-mediated communication*, Hemel Hempstead: Harvester, Ed M. Lea

Star, S. L. and J.R. Griesemer (1989) "Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39." *Social Studies of Science* 19(3), 387-420.

Stavrositu, C.D. and J. Kim (2014) "Social media metrics: Third-person perceptions of health information" *Computers in Human Behavior* 35, 61-67

Stone, D.N. and D.A. Schkade (1991). "Numeric and linguistic information representation in multiattribute choice." *Organizational Behavior and Human Decision Processes* 49 (1), 42-59

Strathern, M. (2000). *Audit Cultures: Anthropological Studies in Accountability, Ethnics and the Academy.* New York: Routledge

Sun, N., Rau, P.P. and L. Ma (2014) "Understanding lurkers in online communities: A literature review" *Computers in Human Behavior* 38, 110-117

Sundararajan, A., Provost, F., Oestreicher-Singer G. and S. Aral (2013) "Information in digital, economic and social networks" *Information Systems Research* 24 (4), 883-905

Sunstein, C. (2003) "What's available? Social influences and behavior economics" *Northwestern University Law Review* 97 (3), 1295-1314

Sweney, M. (2014). *Last.fm made loss of ?2.1m last* year. URL: http://www.theguardian.com/media/2014/oct/08/last-fm-made-loss (visited on 11/21/2014).

Tempini, N. (2015) "Governing PatientsLikeMe: Information production and research through an open, distributed, and data-based social media network" *The Information Society* 31 (2), 193-211

Tilson, D., Lyytinen, K. and C. Sorensen (2010) "Digital Infrastructure: The Missing IS Research Agenda" *Information Systems Research* 21 (4), 748-759

Tirunillai, S. and G. Tellis (2012) "Does chatter matter? The impact of online generated content on a firm's financial performance" *Marketing Science* 31 (2), 198-215

Tong, S.T., Van Der Heide, B. and L. Langwell (2008) "Too much of a good thing? The relationship between number of friends and interpersonal impressions on Facebook" *Journal of Computer-Mediated Communication* 13 (3), 531-549

Tosaka, Y. and C. Weng (2011) "Reexamining content-enriched access: Its effect on usage and discovery." *College & Research Libraries* 72(5), 412-427.

Trattner, C. and F. Kappe (2013) "Social stream marketing on Facebook: a case study" *International Journal of Social and Humanistic Computing* 2 (1-2), 86-103

Van der Vlist (2016) "Accounting for the social: Investigating commensuration and Big Data practices at Facebook" *Big Data & Society* January-June, 1-16

van Dijck, J. (2013) *The culture of connectivity: A critical history of social media.* Oxford: Oxford University Press.

van Dijck, J. (2014) "Datafication, dataism and dataveillance" *Surveillance & Society* 12 (2), 197-208

van Dijck, J. and T. Poell (2013) "Understanding Social Media Logic" *Media and Communication* 1 (1), 2-14

van Dijck, J. and T. Poell (2016) "Understanding the promises of premises of online health platforms" *Big Data & Society* January-June, 2016

Varian, H. R. (2010) "Computer mediated transactions." *American Economic Review* 100(2), 1-10.

Villard, H. and M.A. Moreno (2012) "Fitness on Facebook: Advertisements Generated in Response to Profile Content" *Cyberpsychology, Behavior, and Social Networking* 15 (10), 564-568

Wasserman, S. and K. Faust (1994) *Social Network Analysis: Methods and Applications.* Cambridge University Press.

Wegner, P. (1997) "Why Interaction is More Powerful than Algorithm." *Communications of the ACM* 40 (50), 80-91

Weick, K. (1979). *The Social Psychology of Organizing*. 2nd Edition. USA: McGraw-Hill

Weinberger, D. (2008). *Everything is Miscellaneous: The Power of the New Digital Disorder.* New York: Henry Holt and Company

White, H. C., Boorman, S. A. and R.L. Breiger (1976) "Social structure from multiple networks. I. Blockmodels of roles and positions." *American Journal of Sociology* 81(4), 730-780.

Wilson, R.E., Gosling, S.D., and L.T. Graham (2012) "A Review of Facebook Research in the Social Sciences". *Perspectives on Psychological Science* 7 (3), 203-220

Winiwater, W., Hofferer, M., and B. Knaus. (1997) "CIFS - A Cognitive Information Filtering System with Evolutionary Adaptation." *User Modeling and User-Adapted Interaction.* 7 (3)

Winter, S., Bruckner, C. and N.C. Kramer (2015) "They came, they liked, they commented: Social influence on Facebook News channels" *Cyberpsychology, Behavior and Social Networking* 18 (8), 431-6

Wohn, D.Y. and B.J. Howe (2016) "Micro agenda setters: The effect of social media on young adults' exposure to and attitude toward news" *Social Media + Society* January-March, 1-12

Wohn, D.Y., Can, C.T. and R.A. Hayes (2016) "How affective is a "Like"?: The effect of paralinguistic digital affordances on perceived social support" *Cyberpsychology, Behavior and Social Networking* 19 (9), 562-566

Wu, L., Lin, C., Aral, S. and E. Brynjolfsson (2009) "Network structure and information worker productivity: New evidence from the global consulting services industry" In *Winter Conference Business Intelligence*

Yan, L., Peng, J. and Y. Tan (2015) "Network Dynamics: How Can We Find Patients Like Us?" *Information Systems Research* 26(3), 496-512.

Yang, G. (2016) "Narrative Agency in Hashtag Activism: The Case of #BlackLivesMatter" *Media and Communication* 4 (4), 13-17

Yin, R. K. (2009). *Case Study Research: Design and Methods*. Sage Publications.

Yoo, Y., Henfridsson, O., and K. Lyytinen (2010) "The New Organizing Logic of Digital Innovation: An Agenda for Information Systems research" *Information Systems Research* 21 (4), 724-735

Zappavigna, M. (2011) "Ambient affiliation: A linguistic perspective on Twitter" *New Media & Society* 13 (5), 788-806

Zeng, X. and L. Wei (2013) "Social Ties and User Content Generation: Evidence from Flickr" *Information Systems Research* 24 (1), 71-87

Zervas, P. and D.G. Sampson (2014) "The effect of users' tagging motivation on the enlargement of digital educational resources metadata" *Computers in Human Behavior* 32, 292-300

Zhang et al. (2015a) "Mapping developing of social media research through different disciplines: Collaborative learning in management and computer science" *Computers in Human Behavior* 51, 1142-1153

Zhang et al. (2015b) "From e-learning to social-learning: Mapping development of studies on social media-supported knowledge management" *Computers in Human Behavior* 51, 803-811

Zhang, W. and S.A. Watts (2008) "Capitalizing on content: information adoption in two online communities" *Journal of the Association for Information Systems* 9 (2), 73-94

Zhang, Y. and L. Leung (2015) "A review of social networking server (SNS) research in communication journals from 2006 to 2011" *New Media & Society* 17 (7), 1007-1024

Zimmer, M. (2009) "Renvois of the past, present and future: hyperlinks and the structuring of knowledge from Encyclopedia to Web 2.0" *New Media & Society* 11(1&2), 95-114

Zittrain, J. (2008) *The Future of the Internet and How to Stop it.* New Haven, CT: Yale University Press