Characterizing Information Diets of Social Media Users

Juhi Kulshrestha, Muhammad Bilal Zafar, Lisette Espin Noboa,
Krishna P. Gummadi, Saptarshi Ghosh
Max Planck Institute for Software Systems
Kaiserslautern-Saarbruecken, Germany

Abstract

With the widespread adoption of social media sites like Twitter and Facebook, there has been a shift in the way information is produced and consumed. Earlier, the only producers of information were traditional news organizations, which broadcast the same carefully-edited information to all consumers over mass media channels. Whereas, now, in online social media, any user can be a producer of information, and every user selects which other users she connects to, thereby choosing the information she consumes. Moreover, the personalized recommendations that most social media sites provide also contribute towards the information consumed by individual users. In this work, we define a concept of *information diet* – which is the topical distribution of a given set of information items (e.g., tweets) – to characterize the information produced and consumed by various types of users in the popular Twitter social media. At a high level, we find that (i) popular users mostly produce very specialized diets focusing on only a few topics; in fact, news organizations (e.g., NYTimes) produce much more focused diets on social media as compared to their mass media diets, (ii) most users' consumption diets are primarily focused towards one or two topics of their interest, and (iii) the personalized recommendations provided by Twitter help to mitigate some of the topical imbalances in the users' consumption diets, by adding information on diverse topics apart from the users' primary topics of interest.

Introduction

The rapid adoption of social media sites like Twitter and Facebook is bringing profound changes in the ways information is produced and consumed in our society. Traditionally, people acquired information about world events via *mass media*, i.e., dedicated news organisations that relied on some *broadcast* medium like print (NYTimes or Economist), radio (NPR, BBC radio), or television (CNN, ESPN) to disseminate the information to large numbers of users. Mass media communications are characterised by (i) a small number (few tens to a few hundreds) of news organisations controlling what hundreds of millions of users consume, (ii) an expert team of editors at each news organisation carefully

vetting and selecting news stories to ensure a *balanced coverage* of important news stories, and (iii) all consumers receiving the same standardised information broadcast by each mass media source.

In contrast to the organised world of information production and consumption in broadcast mass media, online social media sites like Twitter and Facebook offer a chaotic information marketplace for millions of producers and consumers of information. Unlike mass media, in social media (i) any of the hundreds of millions of users of these systems can be a producer as well as a consumer of information, (ii) these individual users are *not* expected to provide a balanced coverage of news-stories – they publish any information that they deem important or necessary to share with their friends in real-time, and (iii) information consumption is *personalised* and not all users consume the same information – every individual user selects (e.g., by establishing social links) her preferred sources of information from the millions of individual producers, and *recommender systems* deployed by social media platforms provide an additional source of information to the user. Thus, individual social media users might receive information that is not only unbalanced in terms of coverage of news-stories, but is also very different from what other users in the system receive.

An entire discipline, *media studies*, has largely focused on analysing the coverage of information published on broadcast mass media and how it impacts the consumers of mass media. In contrast, research on understanding the composition of information produced and consumed by social media users is still in its infancy, being limited to a few macroscopic studies on the amounts of information posted by broad categories of users (e.g., celebrities) (Wu et al. 2011; Kwak et al. 2010). There has not been much work on analysing the *composition* of the information produced or consumed by users at the granularity of individual messages.

In this paper, we take the first step towards addressing this challenge by defining the notion of *information diet*. Similar to diet in nutrition, information diet of a user refers to the composition of all the information consumed or produced by the user (Johnson 2012). Specifically, we focus on the *topical composition* of users' diets, i.e., the fraction of their information diets that correspond to different topical categories of information (e.g., information on politics, sports, entertainment, and so on).

One of our key goals is to better understand how the differences in information production and consumption processes between broadcast mass media and online social media affect users' diets. So we conducted a comparative analysis of the topical compositions of the information diets produced, consumed, and recommended on social media and the mass media. Our investigation focused on the following three high-level questions:

1. **Production**: What is the topical composition of information published on broadcast mass media (e.g., NYTimes print edition)? How does the information produced by social media accounts compare with the information published on mass media?

2. **Consumption**: How balanced or unbalanced are consumption diets of social media users (relative to mass media diet)? Are users' consumption diets heavily skewed towards a few topics of their interest, or do they also tend also receive information on a broad variety of topics covered in mass media?

3. **Recommendations**: Do personalised recommender systems deployed by the social media platform provide balanced or unbalanced diets (relative to mass media) to social media users? Do they mitigate or exacerbate the imbalances in the users' consumption diets?

We attempt to address the above questions in the context of the Twitter social media platform. To conduct our study, we needed a methodology to infer the topics of individual posts on Twitter. The bounded length of tweets makes it challenging to infer topics at the level of individual tweets. We propose a novel methodology to infer the topic of a post by leveraging the topical expertise of the Twitter *users* who have posted it. To obtain the information about users' topical expertise, we leverage a methodology based on Twittre Lists, developed in our prior works (Ghosh et al. 2012; Sharma et al. 2012). We show that our methodology performs better at inferring topics for posts than a state-of-the-art publicly deployed commercial topic inference system.

Our study conducted using our above methodology yields several key insights. We highlight a few below:

1. Mass media sources cover a wide range of topics from *politics* and *business* to *entertainment* and *health*. But on social media, the individual sources of information are very focused and publish information dominated by a few topics. It is up to the social media users to select sources to obtain a balanced diet for themselves.

2. We find that for most users, a large fraction of their consumed diet comes from as few as one or two topics, and they hear very little about other niche topics like *health* and *environment* (unless they are interested in these topics).

3. We find that social recommendations, i.e., recommendations about information popular in a user's social network neighbourhood (Gupta et al. 2014), often do not match the user's preferred diet. The differences between recommended and consumed diets are likely due to differences in the interests of a user and the interests of her network

neighbours. As a result, social recommendations introduce topical diversity to a user's diet and can help balance its topical composition.

We have publicly deployed a Web-based service for measuring the information diets produced and consumed by Twitter users, at *http://twitter-app.mpi-sws.org/information-diets/*.

Our work and findings have a number of important implications. As social media becomes more popular, it is important to raise awareness about the balance or imbalance in information diets produced and consumed on social media. Our findings raise the need for better information curators (human editors or automated recommendation systems) on social media that provide a more balanced information diet. Finally, our work is an early attempt, and much future work still remains to be done both on understanding the impact of the diets on consumers in shaping their opinions and on other ways for quantifying the diets beyond topical composition.

Related Work

Analysis of content on mass media: Media studies has been an active field which analyzes the content coverage on mass media, and its effects on the society.[1] There exist a number of 'media watchdog organizations' (e.g., FAIR (*http://fair.org/*), AIM (*http://www.aim.org/*)) which judge the content covered by news organizations based on fairness, balance and accuracy. Additionally, there have also been studies on media biases (Groseclose and Milyo 2005; Budak, Goel, and Rao 2014). Such studies are easier to perform over mass media since it is a broadcast medium and all users receive the same information. On the other hand, studying the information consumed on social media is much more challenging since individual users shape their own personalized channels of information by selecting the other users to follow.

Information production & consumption on social media: Prior studies on information production and consumption on social media (Wu et al. 2011; Kwak et al. 2010; Cha et al. 2012) have been limited to studying the amount of information being exchanged among various users. There has not been any notable effort towards analyzing the *topical composition* of the information produced or consumed, which is the goal of this work.

There have also been some prior works on whether social media users are receiving multiple perspectives on a specific event or topic (Balasubramanyan et al. 2012; Conover et al. 2011; Park et al. 2009; Adamic and Glance 2005; Borge-Holthoefer et al. 2015). Though we focus only on the topical composition of the information produced and consumed by social media users, the concept of information diet introduced in this work can be extended to study opinion polarization on social media.

Topic inference of social media posts: To our knowledge, all prior attempts to infer the topic of a tweet /

---

[1]*http://en.wikipedia.org/wiki/Media_studies*

hashtag / trending topic rely on the content itself – either applying NLP and ML techniques (Quercia, Askham, and Crowcroft 2012; Ramage, Dumais, and Liebling 2010; Ottoni et al. 2014; Zubiaga et al. 2011) or mapping to external sources such as Wikipedia or Web search results (Meij, Weerkamp, and de Rijke 2012; Bernstein et al. 2010) – in order to infer the topics. Such methodologies are of limited utility in the case of social media like Twitter, primarily due to the tweets being too short, and the informal nature of the language used by most users (Sharma et al. 2012; Wagner et al. 2012). In contrast to these previous approaches which focus on the content, our methodology focuses on the characteristics of the *authors* of the content to infer its topic.

Methodology: Quantifying Information Diets

In this paper, we introduce the notion of *information diet* of a set of information items (e.g. a set of tweets or hashtags), as the topical composition of the information items. We define the topical composition over a given set of topics as the fraction of information related to each topic. In this section, we present our methodology for quantifying the information diet for a set of tweets on Twitter.

We chose *hashtags* and *URLs* as the basic elements of information in a tweet and collectively refer to them as *keywords*. However, our methodology can be easily extended to include other kinds of keywords such as named entities. To justify our choice of keywords, we conducted a survey through Amazon Mechanical Turk (AMT: *https://www.mturk.com/*), where we showed workers 500 randomly selected tweets from Twitter's 1% random sample which did *not* contain any keyword. A majority of the AMT workers judged 96% of the tweets without any keywords to be non-topical, i.e., they mostly contained conversational babble. Thus, the hashtags and URLs contain crucial signals about the topicality of tweets, justifying our decision to only consider hashtags and URLs as keywords for inferring the topic of tweets.

The key step in our methodology for quantifying information diets consists of inferring the topic of a keyword, which is described next.

Inferring topic of a keyword

As discussed in the Related Work section, prior approaches for inferring the topic of a tweet / keyword rely on the content itself. Such approaches tend to perform poorly on short posts containing informal language (Sharma et al. 2012; Wagner et al. 2012). So we propose a different technique to infer the topic of a keyword which relies on the topical expertise of the users who are discussing that keyword. The basic intuition behind our technique is that if many users interested in a certain topic are discussing a particular keyword, that keyword is most likely related to that topic.

To identify the topical expertise of users in Twitter, we leveraged the List-based methodology developed in our prior works (Sharma et al. 2012; Ghosh et al. 2012) to retrieve *expertise tags* for topical experts. For instance, some of the tags inferred by this methodology for the expert @ladygaga are 'music', 'entertainment', 'singers', 'celebs' and

Topic categories	Some related terms
Arts-crafts	art, history, geography, theater, crafts, design
Automotive	vehicles, motorsports, bikes, cars
Business-finance	retail, real-estate, marketing, economics
Career	jobs, entrepreneurship, human-resource
Education-books	books, libraries, teachers, school
Entertainment	music, movies, tv, radio, comedy, adult
Environment	climate, energy, disasters, animals
Fashion-style	style, models
Food-drink	food, wine, beer, restaurants, vegan
Health-fitness	disease, mental-health, healthcare
Hobbies	photography, tourism, gardening
Paranormal	astrology, supernatural
Politics-law	politics, law, military, activism
Religion	christianity, islam, hinduism, spiritualism
Science	physics, chemistry, biology, mathematics
Society	charity, LGBT
Sports	football, baseball, basketball, cricket
Technology	mobile-devices, programming, web-systems

Table 1: **The 18 topic-categories to which keywords / tweets will be mapped, and some terms related to each topic. The terms will be matched with expertise-tags.**

'artists'. We extracted topical expertise of 771,000 experts on Twitter by using this methodology. The details of the methodology are omitted here for brevity.

Next, we used two standard topical hierarchies – the Open Directory Project (*www.dmoz.org*) and AlchemyAPI (*www.alchemyapi.com/api/taxonomy/*) – to obtain 18 topical categories and their related terms, as shown in Table 1. The 18 topical categories were selected by combining the top categories of the two hierarchies, while the related terms were derived from their lower levels. In the rest of the paper, we quantify information diets by inferring the fraction of information from each of these 18 topics. We also mapped the experts to one or more of the 18 topic categories, by matching the inferred tags of each expert to the related terms of the topical categories.

As stated earlier, the main intuition behind our methodology is that if several experts on a topic are posting a keyword, then that keyword is most likely related to that topic. To infer the topic of a keyword k, we first identify the set of experts E_k who have posted k. We do *not* attempt to infer the topic of a keyword unless it has been posted by at least 10 of our identified experts. For each topic t (in Table 1), we then determine the fraction (f_t) of experts in E_k who are mapped to that topic t. Next, to account for the varying number of experts mapped to different topics, we normalize the fraction f_t by the total number of experts on topic t in our data set. Finally, we select the topic with the highest normalized fraction f_t to be the inferred topic of keyword k. Further details of the methodology can be found at *http://twitter-app.mpi-sws.org/information-diets/*.

Evaluating the topic inference methodology

We now present the evaluation of the performance of our proposed topic inference methodology, and compare its performance with that of a state-of-the-art commercial service,

Metric	Methodology	Hashtags	
		Popular	Random
Coverage	AlchemyAPI	22.5%	55.5%
	Proposed	98%	82.5%
Accuracy	AlchemyAPI	44.44%	51.35%
	Proposed	58.67%	49.69%

Table 2: **Comparing the proposed topic inference methodology with AlchemyAPI (which uses NLP techniques) in terms of coverage and accuracy.**

AlchemyAPI, that uses NLP and deep-learning techniques for topic inference. We found the performances to be very similar for both hashtags and URLs; hence, for brevity, we only present the evaluation results for hashtags.

The set of hashtags used for evaluation is derived from the Twitter 1% random sample [2] from a week in December 2014. It consists of: (i) 200 popular hashtags which were most tweeted, and (ii) 200 randomly selected hashtags. We inferred the topic of a hashtag using AlchemyAPI by passing 1000 randomly selected tweets containing the hashtag. Table 2 compares the performance of the proposed methodology with AlchemyAPI, based on two metrics - coverage and accuracy.

Coverage: It is defined as the fraction of keywords for which a methodology is able to infer a topic. Table 2 shows that our proposed methodology performs significantly better than AlchemyAPI, which possibly fails due to the informal and abbreviated language used in most tweets. Note that our methodology is able to infer topics for a relatively smaller fraction of random hashtags than the popular ones, since we need the hashtag to be posted by at least 10 experts.

Accuracy: It is defined as fraction of keywords for which the inferred topic is relevant. Relevance was judged through an AMT survey – we showed the hashtag, 20 random tweets containing the hashtag, and the inferred topic to five AMT workers and asked them to judge if the inferred topic of the hashtag is relevant. Table 2 shows the majority opinion of the five workers – the proposed methodology is accurate for a larger fraction of popular hashtags, while AlchemyAPI performs slightly better for randomly selected hashtags.

Overall, our proposed methodology performs better than a state-of-the-art NLP-based technique in inferring topics of hashtags, especially for popular ones – not only does the proposed methodology infer topics for more hashtags, but also the inferred topics are more accurate.

Quantifying information diet of social media posts

Having established the methodology to infer the topic of a keyword, we now use it to construct the information diet of a set of tweets. We first extract the keywords from every tweet in the set and infer the topic of each individual keyword. We then construct a *topic-vector* for the given set of tweets, where the weight of a topic is the total contribution of all keywords inferred to be on that topic. Since a tweet can

---

contain multiple keywords, we normalize the contribution of each keyword within a tweet by the number of keywords in that tweet (so that each tweet contributes a total weight of 1 to the topic-vector). This topic-vector represents the information diet of the given set of tweets.

Limitations of our methodology

We briefly discuss some limitations in our approach of quantifying the information diets of users. First, since we infer the topics of only those keywords which have been tweeted by at least 10 topical experts, we have a lower coverage and accuracy for non-popular keywords. However, the later sections show that the popular information forms a large fraction of users' diets; hence, the approach is likely to be able to estimate the information diets of users fairly accurately.

Second, while we only focus on information that a user posts or consumes on Twitter, we are aware that a user in Twitter is also likely to get information from other online as well as off-line sources. However, as users are relying more and more on social media sites such as Twitter and Facebook to find interesting information (Jane Sasseen et al. 2013), what a user consumes in Twitter is likely to be an increasingly significant factor in shaping her overall information diet.

Mass Media Diet

As mentioned earlier, the goal of this study is to compare and contrast the processes of production and consumption of information over broadcast mass media and over social media. We analyze the information being published over mass media by three popular news organizations – NYTimes, Washington Post and The Economist. We collected their broadcast print editions for three days in December 2014, and categorized the news-articles into our 18 topic-categories (Table 1) through human feedback. Each news-article was shown to five distinct workers recruited through AMT, and the majority verdict was considered as the topic for the news-article.

Table 3 shows the mass media information diets of the three news organizations. We find that all the news organizations tend to focus (i.e., post majority of their news-articles) on a few popular topics – politics, entertainment, and sports for NYTimes and Washington Post, and mainly politics and business-finance for The Economist. However, despite their bias towards these few popular topics, the mass media diets also have a spread over the remaining less popular topics – the 12 least popular topics contribute 25% of the diet for NYTimes and 17% for both Washington Post and Economist.

In the following sections, we use these mass media information diets as a baseline for comparing with various information diets on social media.

Production: Social vs. Mass Media Diets

Traditionally, in mass media, editors of news-organizations are expected to ensure that the news-stream has a balanced coverage across various topics of interest of the subscribers, by following definite guidelines. In contrast, every user-account in social media serves as a producer / source of information, and there are no definite guidelines on the content

Topic	NYTimes	Wash. Post	Economist
Arts-Crafts	4.56%	0.0%	1.85%
Automotive	1.34%	0.0%	0.37%
Business-Finance	7.51%	8.65%	28.04%
Career	0.8%	0.48%	0.74%
Education-Books	1.88%	5.29%	3.32%
Entertainment	12.33%	13.94%	1.48%
Environment	3.49%	0.96%	7.01%
Fashion-Style	0.0%	1.44%	0.0%
Food-Drink	4.83%	6.25%	2.21%
Health-Fitness	6.17%	5.29%	2.95%
Hobbies-Tourism	1.34%	0.0%	0.37%
Paranormal	0.27%	0.0%	0.0%
Politics-Law	29.49%	37.5%	35.06%
Religion	2.14%	0.96%	2.95%
Science	1.34%	0.96%	2.58%
Society	3.75%	6.73%	3.32%
Sports	15.01%	9.62%	1.11%
Technology	3.75%	1.92%	6.64%

Table 3: **Mass media information diets of three news organizations, where the topics of the news-articles were judged by AMT workers (top topics highlighted).**

Social media account	Topic of specialization	Contribution of topic	
		Social media	Mass Media
NYTSports	Sports	66.6%	15.0%
nytimesbusiness	Business	66.1%	7.5%
nytimesbooks	Edu-Books	59.1%	1.9%
EconUS	Business	74.4%	28.0%
EconWhichMBA	Education	37.6%	3.3%
	Business	32.1%	28.0%
PostSports	Sports	88.5%	9.6%
PostHealthSci	Science	34.5%	0.96%
	Health	25.1%	5.3%
WaPoFood	Food	60.3%	6.3%

Table 4: **Examples of topic-specific Twitter accounts of news organisations, along with the contribution of their topics of specialization in their production diet.**

being posted by any account. To analyze the effects of these differences, this section compares various information diets being produced in social media with those of mass media (described in the previous section).

News organizations: Social media vs. mass media

We first address the question: *are there differences between the information diets published by news organizations over mass media and social media?* To answer this question, we collected the tweets posted by the Twitter accounts of the three selected news organizations (NYTimes, Washington Post and The Economist) during December 2014, and generated the information diet produced by these news organizations over social media.[3]

Interestingly, we find that each of the three news organizations has multiple accounts on Twitter. These include one primary account (@nytimes, @washingtonpost and @economist) and several *topic-specific accounts* (e.g., @NYTSports, @EconSciTech, @PostHealthSci) each of which specializes in posting news-stories on a particular topic. Table 4 shows some of the topic-specific accounts of the three news organizations, along with the fraction of their production diet that is on the topic of specialization. It is evident that the topic-specific accounts produce a much larger fraction of their diet on their specific topics of specialization, as compared to the mass media diet of the same news organization.

While the topic-specific accounts of the news organizations have thousands to hundreds of thousands of followers,

a much larger number of users subscribe to the primary accounts. For instance, the primary account @nytimes has 15 million followers, while the topic-specific accounts @NYTSports and @nytimesbusiness have 51K and 567K followers respectively. Since most social media users consume the diet produced by the primary account, we compare the social media diet produced by the primary account with the mass media diet of the same news organization.

Figure 1 compares the information diets produced by the three news organizations over mass media, with those produced by their primary Twitter accounts over social media. We find two main differences between the mass media and social media diets of the same news organization. First, the primary accounts of the news organizations in social media tend to publish less content (as compared to the corresponding mass media diets) on those topics for which there exist topic-specific accounts. For instance, for both NYTimes and Washington Post, topics such as sports and food are covered much lesser in the social media diets than in the corresponding mass media diets. Additionally, both the primary and the topic-specific social media accounts of the news organizations tend to be more specialized in their production by focusing on fewer topics, as compared to their mass media diets. For example, while the mass media diet of Economist focuses on both business and politics, the social media diet of @economist focuses solely on business and publishes far lesser content on politics.

In summary, there is an *unbundling of content on social media* by the news organizations through multiple accounts each specializing on a particular topic. This unbundling would enable users in social media to get focused information on their topics of interest by subscribing to the topic-specific accounts. However, the users who subscribe to only the primary account of the news organizations might not be aware that they are receiving a different information diet as compared to that of the mass media versions.

Popular social media accounts vs. mass media

Next, we study whether our observations about the specialized production of the social media accounts of news organizations generalizes to other popular user-accounts in Twitter.
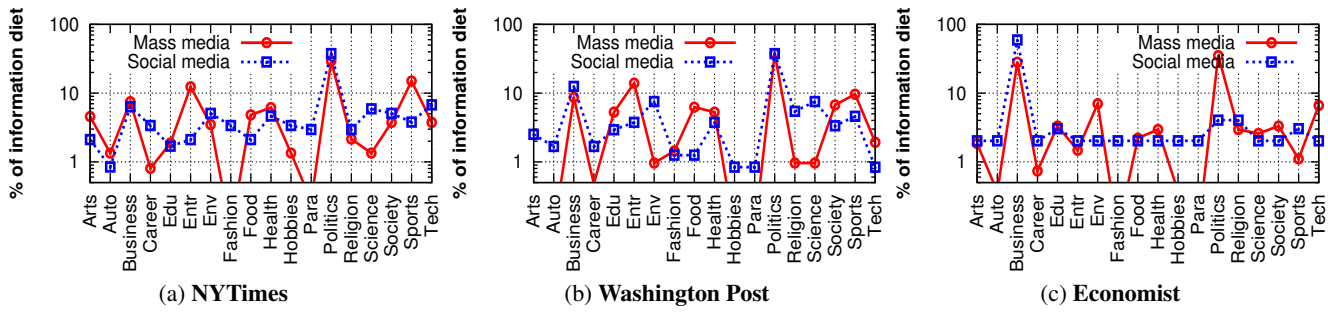
---

[3]The statistics presented in this section are for the same three days in December 2014, over which the mass media diets were analyzed in the previous section. However, we observed that the information diets remain relatively unchanged over longer time-durations.

Figure 1: **Comparing the information diet posted by news organizations in mass media (news-articles in print editions) and social media (tweets posted by the primary Twitter accounts) for the same days in December 2014. Topics with contribution less than 0.5% not shown.**
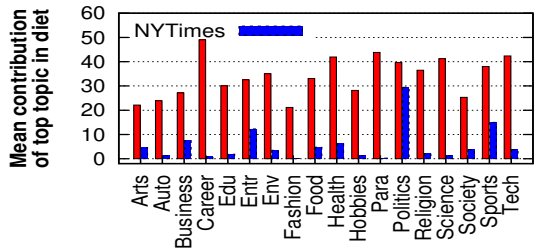

Figure 2: **Mean contribution of the top topic (on which a user posts the highest fraction of her diet) for popular users who focus on different topics.**
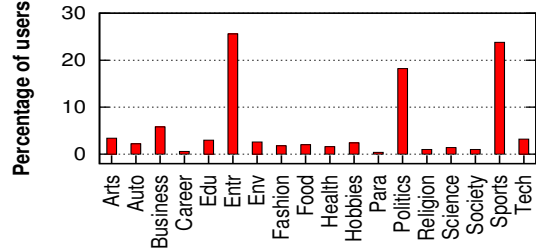

Figure 3: **Distribution of the 500 randomly selected verified users, according to the topic on which they produce the maximum fraction of their diet.**

There are several ways to identify popular / influential accounts in Twitter, such as by the number of followers, or by the number of times one is retweeted. In this study, we consider *verified users* as examples of popular user-accounts on Twitter. Out of all the verified users on Twitter who declared their language as English, and were not news organizations, we randomly selected a set of 500 verified users. We collected the tweets posted by them during December 2014, and computed the information diet posted by these users by the methodology presented earlier.

For studying the specialization in the produced diet of each user, we define the *top topic* for her as the topic on which she posts the largest fraction of her diet. For the group of users having a common top topic, we compute the mean percentage contribution of their posted diet that is on their top topic. Figure 2 shows this mean percentage contribution for the group of users specializing on each topic. As a baseline, we also show the contribution of each topic in the NYTimes mass media diet (which was stated in Table 3). We find that the popular users, on average, post a significant fraction of their diet (between 20% and 50%) on just their top topic. Further, users having different top topics are focused to different degrees – for instance, popular users having career, health, paranormal, science and technology as their top topic post more than 40% of their diet on their top topic. Anyone who subscribes to these popular sources of information on social media will get a much higher frac-

tion of content on the corresponding topic, than what is obtained from a typical mass media source (as shown by the NYTimes baseline in Figure 2).

Additionally, we looked at the distribution of the 500 randomly selected verified users across their top topics. Figure 3 shows the distribution of these users according to their top topic. Most of the users have their top topic as one of the three topics – *entertainment*, *sports* and *politics*. However, there are small fractions of popular users focusing their diets on all the other topics as well. These observations agree with recent findings (Bhattacharya et al. 2014) that though Twitter is primarily thought to be associated with few popular topics such as entertainment, sports, and politics, there are popular accounts who are experts on a wide variety of topics.

These observations imply that, similar to mass media, there are sources of information on a wide variety of topics in the Twitter social media. However, since every source produces a diet that is specialized on just a few topics, the consumers of information in social media need to be careful in deciding whom they subscribe to, especially if they desire to get a topically balanced information diet.

Random sampling of social vs. mass media posts

Till now, we have shown that the individual sources of information in social media (popular user-accounts as well as accounts of news organizations) produce diets that are very
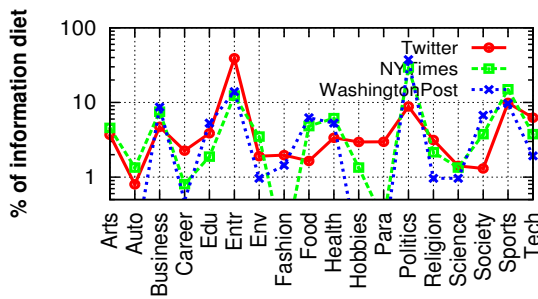
Figure 4: **Comparing the information diet of the Twitter 1% random-sample with the mass media diet of news organizations (NYTimes and Washington Post).**

focused on specific topics. Now we shift the focus to the overall information being produced on the two media. We use the Twitter 1% random sample (for the month of December 2014) to represent the overall information being produced on social media, and compare the information diet of the Twitter random sample with the mass media diets of NYTimes and Washington Post in Figure 4.

We observe that the diets from both social media and mass media are skewed, but towards different topics. Though both diets have entertainment, politics, sports and business amongst the top topics, the Twitter social media diet is more heavily biased towards entertainment (39%), while the mass media diets focus more on politics (30%). Further, some topics are over-represented in the social media diet as compared to mass media diet, such as technology, hobbies-tourism, paranormal, and career. On the other hand, topics such as food, health, and society are covered more in mass media than in social media, which is probably because these topics are of general interest to many people in the off-line world. Whereas, topics such as entertainment and technology are more dynamic, with new information being generated regularly, leading to them being covered more in a real-time information dissemination medium like Twitter.

Consumption: Diets of Social Media Users

Unlike in mass media where everyone consumes the same broadcast information, every user on social media shapes her own personalized channel of consumption by subscribing to other users. In this section, we study how the users are consuming information in social media, as compared to the consumption via mass media.

For this analysis, we selected 500 users randomly from the Twitter userid space (i.e., the user-ids were randomly selected from the range 1 through the id assigned to a newly created account), with the constraint that the selected users follow at least 20 other users (to ensure that the selected users have a meaningful consumption behaviour to study). We then computed the consumed information diet for each user, considering the tweets that a user received from her fol-
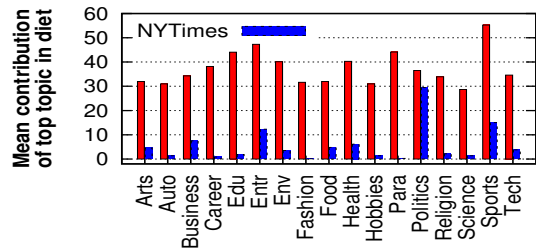

Figure 5: **Mean contribution of the top topic in the consumption diets of random users grouped according to their top topic of consumption.**
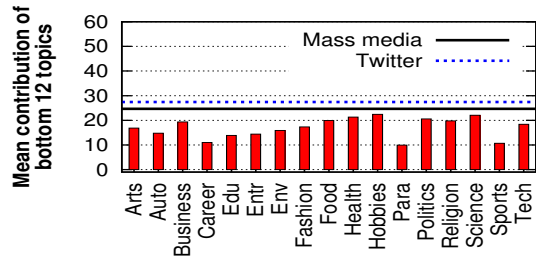

Figure 6: **Mean contribution of the bottom 12 least dominant topics in the consumption diets of random users grouped according to their top topic of consumption.**

lowings (i.e., via word-of-mouth) during the month of December 2014.[4]

Similar to the previous section, we define the top topic for a user as the topic on which she consumes the largest fraction of her diet. For the group of users having a common top topic of consumption, Figure 5 plots the mean contribution of the top topic in the consumption diet of these users.[5] As a baseline, the figure also shows the contribution of each topic in the NYTimes mass media diet. Across almost all topics, the consumers are very focused on their top topic, and on average, consume 30% or more of their diet on that topic. Moreover, when we compute the contribution of up to top two topics, we observe that 80% of the users consume more than half of their diet on only these one or two topics. These observations imply that users in social media consume a much larger fraction of their information diet on their primary topic(s) of interest, as compared to what they would consume on the same topics from a typical mass media source (as shown by the NYTimes mass media baseline).

Additionally, Figure 6 depicts the mean contribution of the *bottom 12 topics* on which the users consume the least information, for the same groups of users. We find that the 'tail topics' account for an inordinately low fraction of their consumed diet. Across all topics, the mean tail topics contribution for users focusing on a particular topic is even lower

---

[4]We consider all tweets received by a user to compute her consumption diet, in the absence of data about what she actually reads.

[5]In our set of 500 randomly selected users, we did not find any user whose top topic of consumption was 'society'; hence we will not consider this topic further in this section.
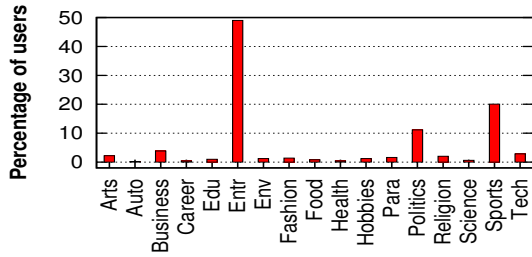
Figure 7: **Distribution of the 500 randomly selected users, according to the topic on which they consume the maximum fraction of their diet.**

than the contribution of the bottom 12 topics in the NYTimes mass media diet (24%) and the Twitter random sample diet (27%).

Finally, Figure 7 plots the distribution of the 500 randomly selected users according to the top topic of consumption. We find that the users' consumptions are very unevenly spread amongst the different topics – as much as half the user population consumes most information on the topic entertainment, while a sizeable fraction focuses on sports and politics. When we compare this distribution to the production distribution of popular users in Figure 3, we observe that consumption behaviours are even more skewed across topics than the production.

Thus we observe that users are extremely selective in the information they consume via social media, with a huge bias towards one or two topics of their interest; moreover, this bias comes at the cost of the tail topics. In future, as users rely more and more on social media like Twitter to consume information, their diets may get progressively more skewed towards the one or two topics of their interest. Users who wish to have a more balanced consumption in social media need to be careful about the sources to which they subscribe. Alternatively, the biases in the consumption diets of users can potentially be mitigated by the information supplied to them by recommender systems deployed in the social media sites; in the next section, we investigate the role of recommender systems in shaping the diets that social media users consume.

Recommendations: Personalisation of Diets

All popular social media systems, including Twitter, deploy recommendation systems to enable users discover content that would be interesting to them. These recommendations expose the users to additional information beyond the information which they get via word-of-mouth over their social follow-links. The recommendation systems currently deployed on most social media largely depend on the (2 hop) social neighborhood of a target user for finding interesting content to recommend to the user (fac ; twi ; Gupta et al. 2014). Hence, such systems are also referred to as *social recommendation systems*.

In the previous section, we saw that the consumed diets of most users are focused on just one or two topics of their interest. In this section, we study the impact of tweet

recommendations on the information that users are exposed to, i.e., whether the recommendations exacerbate or mitigate the topical biases in the consumed diets.

Data collection & methodology

On Twitter, the recommendations provided to a certain user are visible only to her and cannot be crawled publicly. Hence we adopt the methodology of creating *test accounts* on Twitter which *mimic* the followings of real users, i.e., the test accounts have the same network neighborhood as the mimicked real users. We randomly selected 15 real users with their number of followings varying between 10 and 1000 (to ensure that these users have social neighborhoods of different sizes), and created test accounts mimicking these users.[6] We refer to these test accounts as u1, u2, ..., u15.

The recommendations given in Twitter are dynamic, and are updated in real-time (Gupta et al. 2014). Hence, for each test account, we gathered a snapshot of the recommendations every 30 minutes, for a week in December 2014. On an average, each user received 708 recommended tweets in each gathered snapshot. Since these are too many for any user to view practically, we considered only the top 10 recommended tweets per snapshot.[7] We also collected the tweets received by each test account from all her followings, during the same period in December 2014.

For each of the 15 test accounts, we construct 3 information diets: (i) *consumed diet*: the tweets it receives directly from the users it is following, (ii) *recommended diet*: the top tweets recommended to it, and (iii) *combined diet*: assuming that the user pays equal attention to the consumed and the recommended diets, this is constructed by considering the average contribution from consumed and recommended diets for each topic.

Recommended diets vs. Consumed diets

We first investigate *whether the recommendations are personalized for each user*, i.e., whether different users get different recommended diets. Table 5 states the variation (range) in the percentage contribution of some of the topics in the recommended diets given to the 15 test accounts. It is evident that different accounts are being recommended different diets, with varying contributions of topics.

Next, we examine *the extent to which the recommendations given to a certain user match the consumed diet of the user*. In other words, assuming that the top topics in the consumed diet reflect the topical interests of the user, does the recommended diet contain more or less of the same topics?

To quantify how well the recommended diet matches the consumed diet of a user, we use the standard measure of KL-divergence (KLdiv in short) of the recommended diet

---

[6]These are passive test accounts which do *not* perform any activity such as tweeting or favoriting, etc. They only gather the recommendations given to them by Twitter. Even though the creation of such test accounts results in some users gaining an extra follower, we believe that this has negligible effect on a large social network like Twitter.

[7]We verified that the insights presented later in the section hold even if we consider all recommended tweets (instead of the top 10).

Topic	Range (%)	Topic	Range (%)
Automotive	0.59 – 10.83	Business	2.01 – 18.01
Entertainment	**5.14 – 40.36**	Environment	1.27 – 6.11
Food	0.49 – 4.32	Health	0.79 – 5.45
Politics	**9.03 – 33.34**	Religion	1.76 – 6.81
Science	3.57 – 13.05	**Sports**	**6.14 – 46.97**

Table 5: **Range of contributions of different topics in the recommended diets given to the test accounts.**

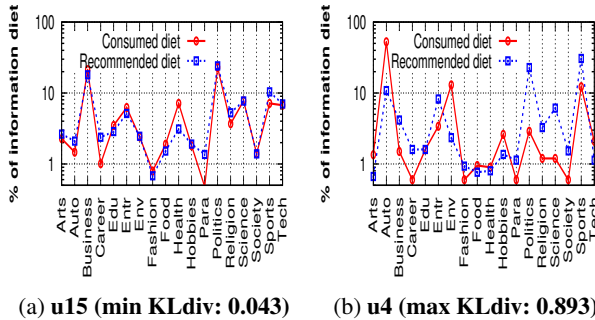(a) **u15 (min KLdiv: 0.043)**   (b) **u4 (max KLdiv: 0.893)**

Figure 8: **Comparing the consumed diet and recommended diet of two test accounts – (i) the one with the minimum KL divergence, and (ii) the one with the maximum KL divergence of the recommended diet from the consumed diet.**

(a) **Top 3 topics**   (b) **Bottom 12 topics**

Figure 9: **Contribution of the (i) top 3 consumed topics and (ii) bottom 12 consumed topics in the consumed, recommended and combined diets for the test accounts.**

tion is popular in the social neighborhood of the user (Gupta et al. 2014).

The effect of the social neighborhood can also be observed from Table 5 where it is seen that popular topics like entertainment, politics and sports are being recommended to everyone irrespective of whether they are interested in these topics. Every account is getting recommended at least 5%, 9% and 6% in entertainment, politics and sports respectively, which is significantly higher than for other topics. On an average, every test account receives up to 17%, 19% and 17% on entertainment, politics and sports respectively. As observed in earlier sections, there are a large number of users tweeting about these topics of general interest (see Figure 3), and hence everyone's neighborhood is likely to contain significant discussions on these topics, which get included into the social recommendations.

Comparing with mass media diet

Finally, we address the question *whether the recommendations mitigate or exacerbate the biases in the users' consumed diets*. For this, we consider the top 3 topics in the consumed diet of an account (i.e., the 3 topics on which the account consumes most information from its followings), and measure the contribution of these 3 topics in the consumed, recommended and combined diets of the user. These are plotted for the 15 test accounts in Fig. 9(a). Similarly, the Fig. 9(b) shows the contribution of the bottom 12 topics in the consumed diet of an account in the three diets.

Interestingly, we observe that the top 3 consumed topics account for a significantly smaller share in the recommended diets of the users, as compared to the consumed diets. As a result, the combined diets of the users also contain a lesser contribution from these three topics, as compared to the consumed diets. Again, the contribution of the bottom 12 topics is higher for the recommended and combined diets, as compared to the consumed diets of the users. Thus, the recommendations tend to even out the imbalances in the consumed diets of the users, by including information from the lower ranking topics in user's consumed diets. Hence, social recommendations are reducing the gap between the information that different users are exposed to by mitigating the biases in the user's diets. To quantify this mitigation, we computed the

from the consumed diet. The smaller the value of KLdiv, the closer the two diets are. We observe that the KLdiv values for the 15 test accounts vary in the range of 0.043 to 0.893, with 5 accounts having KLdiv values below 0.2, and 3 having values above 0.4. This variation in the KLdiv values suggests that the recommendations match the consumed diets to different extents for different users.

Figure 8 shows the topical compositions of the consumed and recommended diets for two test accounts – (i) u15 which has the minimum KLdiv, and (ii) u4 which has the maximum KLdiv of the recommended diets from their consumed diets. It can be seen that the recommended diet of u15 largely matches the consumed diet, while for u4 there is greater mismatch between the two diets. For instance, though u4 consumes a lot of information on the topics automobile and environment, its recommended diet has much lower fraction of these topics. On the other hand, the recommended diet for u4 has higher fractions of politics, religion, and science, topics which are not that significant in its consumed diet.

These observations suggest that the recommended diet that a user will get, does not always match her consumed diet. We also observe cases where two accounts are consuming approximately the same amount of information on a particular topic, but they receive very different amounts of information on this topic in their recommended diets. These differences may be driven by the fact that different users have different social neighborhoods, and the social recommendations given by Twitter are derived from what informa-
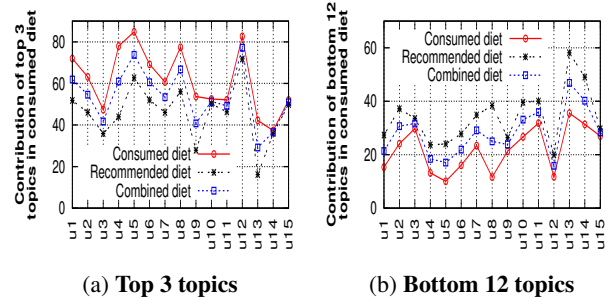
KL-divergence between a user's (i) consumed and (ii) combined diets, from the baseline of the NYTimes mass media diet. We found that, for each of the accounts, the divergence from the baseline is lesser for the combined diet than for the consumed diet, showing that the social recommendations are actually having an equalizing effect across the users (and driving the combined diets towards the baseline).

Thus, we find that social recommendations mitigate the imbalances in the users' consumed diets, bringing in more heterogeneity into what the users are being exposed to.

Concluding Discussion

In this work, we introduced the concept of *information diet* which is the topical composition of the information that is consumed or produced by a user. We proposed a novel methodology for quantifying information diets, by inferring the topics of tweets and keywords in the Twitter social media. Our findings show that (i) individual information sources (user-accounts) on social media produce information that is very focused on a few topics, (ii) most users consume information primarily on one or two topics, and are often not careful about shaping a balanced diet for themselves, and (iii) social recommendations somewhat mitigate the imbalances in the users' consumed diets by adding some topical diversity.

We envisage that this work will not only create awareness among social media users about potential imbalances in their information diets, but will also have implications for the designers of future information discovery, curation and recommendation systems for social media. For instance, we found that social recommender systems are bringing in more heterogeneity into what the users are being exposed to. While this is good for broadening the horizons for the users, topic-specific recommendations might be necessary to provide information focused on the users' interests. Studying the information diets provided by different types of recommender systems, and their impact on the information that a user is exposed to, is an interesting direction to pursue in the future.

References

Adamic, L., and Glance, N. 2005. The political blogosphere and the 2004 US election: divided they blog. In *Proc. LinkKDD*.

Balasubramanyan, R.; Cohen, W. W.; Pierce, D.; and Redlawsk, D. P. 2012. Modeling Polarizing Topics: When Do Different Political Communities Respond Differently to the Same News? In *Proc. ICWSM*.

Bernstein, M. S.; Suh, B.; Hong, L.; Chen, J.; Kairam, S.; and Chi, E. H. 2010. Eddi: interactive topic-based browsing of social status streams. In *Proc. ACM UIST*.

Bhattacharya, P.; Ghosh, S.; Kulshrestha, J.; Mondal, M.; Zafar, M. B.; Ganguly, N.; and Gummadi, K. P. 2014. Deep Twitter Diving: Exploring Topical Groups in Microblogs at Scale. In *Proc. ACM CSCW*.

Borge-Holthoefer, J.; Magdy, W.; Darwish, K.; and Weber, I. 2015. Content and Network Dynamics Behind Egyptian Political Polarization on Twitter. In *Proc. ACM CSCW*.

Budak, C.; Goel, S.; and Rao, J. M. 2014. Fair and Balanced? Quantifying Media Bias Through Crowdsourced Content Analysis. *http://dx.doi.org/10.2139/ssrn.2526461*.

Cha, M.; Benevenuto, F.; Haddadi, H.; and Gummadi, K. 2012. The world of connections and information flow in twitter. *IEEE Trans. Systems, Man and Cybernetics - Part A* 42:991–998.

Conover, M. D.; Ratkiewicz, J.; Francisco, M.; Goncalves, B.; Flammini, A.; and Menczer, F. 2011. Political polarization on Twitter. In *Proc. ICWSM*.

The Facebook Blog — People You May Know. *https://blog.facebook.com/blog.php?post=15610312130*.

Ghosh, S.; Sharma, N.; Benevenuto, F.; Ganguly, N.; and Gummadi, K. 2012. Cognos: Crowdsourcing Search for Topic Experts in Microblogs. In *Proc. ACM SIGIR*.

Groseclose, T., and Milyo, J. 2005. A measure of media bias. *The Quarterly Journal of Economics* 120:1191–1237.

Gupta, P.; Satuluri, V.; Grewal, A.; Gurumurthy, S.; Zhabiuk, V.; Li, Q.; and Lin, J. 2014. Real-Time Twitter Recommendation: Online Motif Detection in Large Dynamic Graphs. In *Proc. VLDB*.

Jane Sasseen et al. 2013. Digital: As Mobile Grows Rapidly, the Pressures on News Intensify — State of the Media (Pew Research Center). *http://tinyurl.com/state-of-media/*.

Johnson, C. A. 2012. *The Information Diet: A Case for Conscious Consumption*. O'Reilly Media.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is Twitter, a social network or a news media? In *Proc. WWW*.

Meij, E.; Weerkamp, W.; and de Rijke, M. 2012. Adding semantics to microblog posts. In *ACM WSDM*.

Ottoni, R.; Las Casas, D.; Pesce, J. P.; Meira Jr, W.; Wilson, C.; Mislove, A.; and Almeida, V. 2014. Of pins and tweets: Investigating how users behave across image-and text-based social networks. In *Proc. ICWSM*.

Park, S.; Kang, S.; Chung, S.; and Song, J. 2009. NewsCube: delivering multiple aspects of news to mitigate media bias. In *Proc. ACM CHI*.

Quercia, D.; Askham, H.; and Crowcroft, J. 2012. Tweetlda: supervised topic classification and link prediction in twitter. In *Proc. ACM Web Science Conference*.

Ramage, D.; Dumais, S.; and Liebling, D. 2010. Characterizing microblogs with topic models. In *Proc. ICWSM*.

Sharma, N.; Ghosh, S.; Benevenuto, F.; Ganguly, N.; and Gummadi, K. 2012. Inferring Who-is-Who in the Twitter Social Network. In *Proc. ACM WOSN*.

Twitter Blogs — Discover: Improved personalization algorithms and real-time indexing. *http://tinyurl.com/twitter-discover-tab*.

Wagner, C.; Liao, V.; Pirolli, P.; Nelson, L.; and Strohmaier, M. 2012. It's not in their tweets: modeling topical expertise of Twitter users. In *Proc. IEEE SocialCom*.

Wu, S.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Who says what to whom on Twitter. In *Proc. WWW*.

Zubiaga, A.; Spina, D.; Fresno, V.; and Martínez, R. 2011. Classifying trending topics: a typology of conversation triggers on twitter. In *Proc. ACM CIKM*.