Characterizing Information Diets of Social Media Users

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 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 two to three 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, (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 (med.). However, research on understanding the composition of information produced and consumed by social media users is still in its infancy. There have been a few macroscopic studies examining whether Twitter is a social network or a news media (Kwak et al. 2010) and the relative amounts of information posted by broad categories of users (e.g., celebrities) (Wu et al. 2011), but there has not been much work on analysing the nature (e.g., the topical 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 introducing the concept of information diet (Johnson 2012). Similar to diet in nutrition, information diet of a user refers to the topical composition of all the information consumed or produced by the user. 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).\(^1\)

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 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 and interests of the Twitter users who have posted it. We obtained information about users’ topical interests from prior work (Ghosh et al. 2012; Sharma et al. 2012; Wagner and others 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 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 1 or 2 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, 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 her network neighbours. As a result, social recommendations introduce topical diversity to a user’s diet and help balance its topical composition.

Our work and findings have a number of important implications. First, 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 represents 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 to quantify the diets (beyond topical decomposition).

**Related Work**

**Analysis of content on mass media**: Media studies (med ) has been an active field which analyzes the content coverage on mass media, and its effects on the society. There exist a number of ‘media watchdog organisations’ (e.g., FAIR (http://fair.org/), AIM (http://www.aim.org/)) which judge the content covered by news organisations 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.

On the other hand, we study the imbalances in the topical composition of information produced and consumed over social media, where not only are all the users capable of producing information, but the users themselves shape individualised channels of information (different from other users) by selecting the other users to follow. Note that this study focuses on topical composition of information on social media, and not on presence of different perspectives (or biases) within a topic. However, the concept of information diet introduced in this work can be extended to study opinion polarization on social media.

**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. In this paper, we address this challenge by the introducing a concept of information diet, which gives interesting insights about the exchange of information on specific topics.

**Topic inference of social media posts**: To our

\(^1\)Note that our focus in this work in on analysing the different topics on which a user produces and consumes information, but not on whether the user is receiving multiple perspectives on a specific event within a topic (e.g., Ferguson protests in US politics). There have been some prior studies that focused on the latter goal (Adamic and Glance 2005; Balasubramanyan et al. 2012; Borge-Holthoefer et al. 2014; Conover et al. 2011; Park et al. 2009).
knowledge, all prior attempts to infer the topic of a tweet / 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; Lee 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; Michelson and Macskassy 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 and others 2012). Contrary to these previous approaches, our methodology instead relies on the characteristics of the authors of a tweet or hashtag, to infer its topic.

Methodology: Quantifying Information Diets

We introduce the concept information diet of a set of information items (e.g. a set of tweets or hashtags) which is the topical composition of the information-set. The topical composition is defined 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. 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 and just contain conversational babble. This justifies 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 and this is described next.

Inferring topic of a keyword

As stated in the Related Work section, prior approaches for inferring the topic of a tweet / keyword rely on the content itself. Since such approaches do not perform well on short posts containing informal language (Sharma et al. 2012; Wagner and others 2012), we propose a completely different technique to infer the topic of a keyword which relies on the topical characteristics of the users who are discussing that keyword. The basic intuition behind our technique is that if many users related to a certain topic are discussing a particular keyword, that keyword is most likely related to that topic.

To identify topical characteristics of users in Twitter, we leveraged a List-based methodology developed in (Sharma et al. 2012; Ghosh et al. 2012) to identify topical experts in Twitter, and some expertise tags for each expert. For instance, some of the tags inferred by this methodology for the topical expert @ladygaga are ‘music’, ‘entertainment’, ‘singers’, ‘celebs’ and ‘artists’. We extracted 771,000 topical experts on Twitter by using this methodology.

### Table 1: The 18 topic-categories to which keywords (e.g., hashtags, URLs) will be mapped, and some terms related to each topic, as identified from two standard topical hierarchies – ODP and AlchemyAPI.

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

Next, we referred to two standard topical hierarchies – the Open Directory Project (ODP: www.dmoz.org) and AlchemyAPI (http://www.alchemyapi.com/) – to identify 18 topical categories and their related terms 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 would be constructing information diets which are basically the fraction of information from each of these 18 topics. We also mapped the topical experts on Twitter 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 we stated earlier, the main intuition behind our methodology is that if several users related to a topic are posting a keyword, then that keyword is most likely related to that topic. Therefore, to infer the topic of a keyword $k$, we first identify the set of experts $E_k$ who have posted $k$. For each topic $t$, 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 normalise 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 normalised fraction $f_t$ to be the inferred topic of keyword $k$.

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, AlchemyAPI, that uses NLP and deep-learning techniques for topic inference.

2The AlchemyAPI topical hierarchy is available at http://www.alchemyapi.com/api/taxonomy/.

3We do not attempt to infer the topic of a keyword unless it has been posted by at least 10 of our identified topical experts.
Table 2: Comparing the proposed topic inference methodology with AlchemyAPI in terms of coverage and accuracy. Results reported for 200 most popular hashtags and 200 randomly selected hashtags chosen from the Twitter 1% random-sample.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Methodology</th>
<th>Hashtags</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Popular</td>
</tr>
<tr>
<td>Coverage</td>
<td>AlchemyAPI</td>
<td>22.5%</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>98%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>AlchemyAPI</td>
<td>44.44%</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>58.67%</td>
</tr>
</tbody>
</table>

We found the performance results to be very similar for both hashtags & URLs and for brevity, we only present the evaluation results for hashtags here. Our evaluation set is derived from the Twitter 1% random sample 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 the hashtag using AlchemyAPI, by passing it random 1000 tweets containing the hashtag. We also inferred the topic of the keyword by our methodology. We evaluate the performance of the two methodologies based on two metrics - coverage and accuracy, as shown in Table 2.

**Coverage:** It is defined as the fraction of keywords for which the methodology is able to infer a topic. As can be seen from Table 2 the proposed methodology performs significantly better than AlchemyAPI, which possibly fails due to the informal and abbreviated language used in most tweets. On the other hand, our methodology is able to infer the topics for relatively lesser fraction of random hashtags than the popular ones, mostly because 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. We showed the hashtag, 20 random tweets containing the hashtag and the inferred topic to AMT workers and asked them to judge if the inferred topic of the hashtag is relevant. We considered the majority opinion of 5 workers on whether the topic inference was correct. Again the Table 2, shows that 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 topic \( t \) is the total contribution of all keywords inferred to be on that topic \( t \). Since a tweet can contain multiple keywords, we normalise the contribution of each keyword within a tweet by the number of keywords in that tweet. 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, we can infer the topics of only those keywords which have been tweeted by at least 10 topical experts. This leads to us having a higher coverage and accuracy for popular keywords. However, we show in later sections that such 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, though we may miss out on some of the less popular information.

Second, while we only focus on information that a user posts or consumes on Twitter, we are aware that a user is also likely to get information from other online as well as off-line sources. However, prior research shows that users are relying more and more on social media sites such as Twitter and Facebook to find interesting news and information (Sasseen, Jane and Olmstead, Kenny and Mitchell, Amy (Pew Research Center) hence, what a user consumes in Twitter is likely to be an increasingly significant factor in shaping her overall information diet.

**Mass Media Diet**

As stated 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 consider 3 popular news organisations – NY-
Table 4: Examples of topic-specific Twitter accounts of news organisations, along with their topics of specialization. The diets posted by such accounts contain a much higher fraction on their topic of specialization, as compared to the mass media diets of the same news organizations.

<table>
<thead>
<tr>
<th>Social media account</th>
<th>Topic of specialization</th>
<th>Contribution of topic to diet</th>
<th>Social media</th>
<th>Mass Media</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYTSports</td>
<td>Sports</td>
<td>66.6%</td>
<td>15.0%</td>
<td></td>
</tr>
<tr>
<td>nytimesbusiness</td>
<td>Business</td>
<td>66.1%</td>
<td>7.5%</td>
<td></td>
</tr>
<tr>
<td>nytimesbooks</td>
<td>Edu-Books</td>
<td>59.1%</td>
<td>1.9%</td>
<td></td>
</tr>
<tr>
<td>EconUS</td>
<td>Business</td>
<td>74.4%</td>
<td>28.0%</td>
<td></td>
</tr>
<tr>
<td>EconWhichMBA</td>
<td>Education</td>
<td>37.6%</td>
<td>3.3%</td>
<td></td>
</tr>
<tr>
<td>PostSports</td>
<td>Sports</td>
<td>88.5%</td>
<td>9.6%</td>
<td></td>
</tr>
<tr>
<td>PostHealthSci</td>
<td>Science</td>
<td>34.5%</td>
<td>0.96%</td>
<td></td>
</tr>
<tr>
<td>PostWashington</td>
<td>Health</td>
<td>25.1%</td>
<td>5.3%</td>
<td></td>
</tr>
<tr>
<td>WaPoFood</td>
<td>Food</td>
<td>60.3%</td>
<td>6.3%</td>
<td></td>
</tr>
</tbody>
</table>

Times, Washington Post and The Economist – to represent the information being published over mass media. We collected their broadcast print editions for three days in December 2014, and categorised 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 organisations. We find that all the news organisations focus (i.e., post most of their news-articles) on a few specific topics – politics, entertainment, and sports for NYTimes and Washington Post, and mainly politics and business-finance for The Economist. However, in spite of their bias towards these popular topics, the mass media diets also cover non-negligible amounts on most other lesser popular topics as well – 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 shall 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, a lot of attention is paid to the content being posted by news organizations on mass media, where a team of editors attempt to ensure that the news-stream being broadcast is balanced across various topics of interest of the subscribers. In contrast, every user-account in social media serves as a producer / source of information, and there are no definite guidelines on the content 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 organisations: 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 organisations (NYTimes, Washington Post and The Economist) during December 2014, and generated the information diet produced by these news organisations over social media.\(^5\)

Note that each of the three news organisations has multiple accounts on Twitter. These include one primary account (@nytimes, @washingtonpost and @economist) and several topic-specific accounts (such as @NYTSports, @EconSciTech, @PostHealthSci, and so on) 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 organisations, 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 organisations over mass media, with those produced by their primary Twitter accounts over social media. We find that some topics get almost equal focus in both diets, such as politics for NYTimes and Washington Post, and business for Economist. However, there are also some interesting differences between the mass media and social media diets of the same news organization. For instance, for both NYTimes and Washington Post, topics such as entertainment, sports, and food are covered much lesser in the social media diets than in the corresponding mass media diets; on the other hand, few topics such as science and career get covered more in the social media diets. Similarly, 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. These differences suggest that 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.

Thus, there is an unbundling of content on social media by the news organizations through multiple accounts each specializing on a particular topic. This 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

---

\(^5\)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.
they are receiving a different information diet as compared to that of the mass media versions.

**Popular social media accounts vs. mass media**

Prior research has shown that a very large fraction of the information being consumed by users on social media sites like Twitter is produced by a small fraction of popular users (Wu et al. 2011). Hence, we next study the information diet of the content posted by some popular accounts in Twitter.

There are several ways of identifying 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 and were not news organisations, 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.

We find that almost each of the verified users posts a very large fraction of their diet on a particular topic (we refer to this topic as the ‘top topic’ for a user), and this topic varies from user to user. Figure 2 shows the distribution of the 500 randomly selected verified users according to their top topic on which they post the highest fraction of information. 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 (Bhattacharyya 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.

We next study the extent to which individual users focus their posted diet on a specific topic. For the group of popular users having a common top topic (i.e., the groups shown in Figure 2), we compute the mean percentage contribution of their posted diet that is on their top topic. Figure 3 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). It is seen that the popular users, on average, post as much as 20% – 40% of their diets on 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 fraction of content on the corresponding topic, than what is obtained from a typical mass media source (as shown by the NYTimes baseline in Figure 3).

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 one or two 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 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, 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 diets than in the social media.

The above differences can possibly be explained as follows. The topics which dominate the social media (Twitter) diet are generally the ones which frequently have related events taking place – such as sports matches, release of new movies/music albums, political events – and thus new information being generated on a regular basis. Twitter, being a real-time information dissemination medium, forms the perfect substrate for these dynamic topics. On the other hand, the mass media diets contain larger contribution from topics which are of general interest to many people in the off-line world, even though they are not as dynamic, such as food, health and society. Understanding these differences is important for the consumers who need to be aware of the implications of subscribing to the different media.

**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 personalised 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 userids 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 followings (i.e., via word-of-mouth) during the month of December 2014.

Similar to the production diets of popular users we find that consumption of social media users tend to be focused on one or two topics. Specifically, for 80% of the users, more than half of their entire consumed information is contributed by only one or two topics.

Figure 5 plots the distribution of the 500 randomly selected users according to the topic on which they consume the maximum fraction of their diet (i.e., the topic on which they receive the most information from their followings).

![Figure 4: Comparing the information diet of the Twitter 1% random-sample with the mass media diet of news organizations (NYTimes and Washington Post).](image)

![Figure 5: Distribution of the 500 randomly selected users, according to the topic on which they consume the maximum fraction of their diet (i.e., the topic on which they receive the most information from their followings).](image)

6Note that we consider all tweets received by a user while computing her consumption diet, in the absence of data about what a user actually reads.

7In our set of 500 randomly selected users, we did not find any user whose most dominant topic of consumption was ‘society’; hence we will not consider this topic further in this section.
as compared to what they would consume on the same
topic from a typical mass media source (as shown by the
NYTimes mass media baseline).

Additionally, Figure 7 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 than the contribution of the bottom 12 topics in the
NYTimes mass media diet (27%) and the Twitter random
sample diet (24%).

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 cautious 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; the next section investigates the role of
recommender systems in shaping the diets that social media
users consume.

Recommendations: Personalisation of Diets
All popular OSNs deploy recommendation systems to
enable users discover content that would be interesting
to them. These recommendations expose the users to
additional information on top of their consumed diets which
they get via word-of-mouth. The recommendation systems
presently deployed on most social media largely depend
upon the (2 hop) social neighbourhood of a target user
for finding interesting content to recommend to her (fac :
twi ; Gupta and others 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 on which
they consume most information. In this section, we study
the impact of recommendations on the information that
users are exposed to, and whether the recommendations ex-
acerbate 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 real users by having the same
social neighbourhood as them. We randomly selected 15
real candidate users to be mimicked, with the number of
followings varying between 10 and 1000 (to ensure that
these users have social neighborhoods of different sizes).

The recommendations given in Twitter are dynamic,
and are updated in real-time (Gupta and others 2014). Hence,
for each test account, we gathered a snapshot the
recommendations every 30 mins for a week in December
2014. On an average, each user received 708 recommended
tweets every 30 mins 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. However, we verified that the
statistics reported in this section are considering the
average contribution from consumed and recommended
diets for each topic.

Recommended diets vs. Consumed diets
We first investigate whether the recommendations are
personalised for each user, such that different users get

---

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

<table>
<thead>
<tr>
<th>Topic</th>
<th>Range (%)</th>
<th>Topic</th>
<th>Range (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automotive</td>
<td>9.03 – 33.34</td>
<td>Business</td>
<td>2.01 – 18.01</td>
</tr>
<tr>
<td>Entertainment</td>
<td>5.14 – 40.36</td>
<td>Environment</td>
<td>1.76 – 6.81</td>
</tr>
<tr>
<td>Food</td>
<td>0.49 – 4.32</td>
<td>Health</td>
<td>0.79 – 5.45</td>
</tr>
<tr>
<td>Politics</td>
<td>9.03 – 33.34</td>
<td>Religion</td>
<td>1.76 – 6.81</td>
</tr>
<tr>
<td>Science</td>
<td>3.57 – 13.05</td>
<td>Sports</td>
<td>6.14 – 46.97</td>
</tr>
</tbody>
</table>

---

8The statistics reported in this section are considering the
top 10 recommended tweets. However, we verified that the
insights presented later in the section hold even if we consider all
recommended tweets.
The next question which naturally arises is whether the recommendations given to a certain user mimic the consumed diet of the user, and to what extent. In other words, assuming that the top topics in the consumed diet reflect the topical composition of the user, does the recommended diet contain more of the same topics?

To quantify how well the recommended diet matches the consumed diet of a user, we use the standard metric of Kullback–Leibler divergence\(^9\) (KL-div in short) of the recommended diet from the consumed diet. The smaller the value of KL-div, the closer the two diets are. We observe that the KL-div values for the 15 test accounts vary in the range of 0.043 to 0.831, with 5 accounts have KL-div values below 0.2, and 3 having values above 0.4. This suggests that the recommendations mimic the consumed diets to different extents for different users.

Figure 8 shows the topical compositions of the consumed and recommended diets for 2 test accounts; (i) u15 which has the minimum KL-div, and (ii) u4 which has the maximum KL-div of the recommended diets from their consumed diets. It can be seen that the recommended diet of u15 largely mimics 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 there might be some unpredictability in the recommended diet that a user will get, which might not always mimic the consumed diet. We also notice cases where two accounts are consuming approximately the same amount of information on a particular topic, but they receive very different fractions on this topic in their recommended diets. Such unpredictable behaviour may be a side-effect of the fact that the recommendations provided by Twitter are social recommendations (Gupta and others 2014), and hence the topical composition of the recommended diets may vary depending on the availability of information in the social neighbourhood of the target user.

The effect of the social neighbourhood 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 2), and hence everyone’s neighbourhood 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. This goes to show that the recommendations actually 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, we computed the KL-divergence between a user’s (i) consumed and (ii) combined diets, from the baseline of the NYTtimes mass media diet.

---

\(^9\)http://en.wikipedia.org/wiki/Kullback-Leibler_divergence
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 equalising effect across the users (and driving the combined diets towards the baseline).

Hence, we surprisingly find that social recommender systems are actually behaving quite differently from topical recommender systems (where content is recommended on the exact topics of interest of the user). They are in fact bringing in more heterogeneity into what the users are being exposed to. While this is good for broadening the horizons for the users, topical recommendations might be necessary to get recommendations focused on the user’s interests.

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 diet, but will also have implications for the designers of future information discovery, curation and recommendation systems for social media.

References


