Andreas Jungherr, and Pascal Jürgens. 2013. "Forecasting the pulse: how deviations from regular patterns in online data can identify offline phenomena." *Internet Research* 23(5): 589-607. doi: 10.1108/IntR-06-2012-0115

Preprint: Please cite published version!

Forecasting the pulse: how deviations from regular patterns in online data can identify offline phenomena

Abstract

Purpose

The steady increase of data on human behavior collected online holds significant research potential for social scientists. We add to this research by a systematic discussion of different online services, their data generating processes, the offline phenomena connected to these data, and by demonstrating, in a proof of concept, a new approach for the detection of extraordinary offline phenomena by the analysis of online data.

Design/methodology/approach

To detect traces of extraordinary offline phenomena in online data, we determine the normal state of the respective communication environment by measuring the regular dynamics of specific variables in data documenting user behavior online. In our proof of concept, we do so by concentrating on the diversity of hashtags used on Twitter during a given time span. We then use the seasonal trend decomposition procedure based on loess (STL) to determine large deviations between the state of the system as forecasted by our model and the empirical data. We take these deviations as indicators for extraordinary events, which led users to deviate from their regular usage patterns.

Findings

We show in our proof of concept that this method is able to detect deviations in the data and that these deviations are clearly linked to changes in user behavior triggered by offline events.

Originality/value

This paper adds to the literature on the link between online data and offline phenomena. It proposes a new theoretical approach to the empirical analysis of online data as indicators of offline phenomena. The paper will be of interest to social scientists and computer scientists working in the field.

Keywords

Computational social science; Social media; Twitter; Forecasting; Event detection; Diversity; Seasonal trend decomposition procedure based on loess (STL).

Category

Research paper

Andreas Jungherr is a research fellow at the chair of political sociology at the Otto-Friedrich-Universität, Bamberg, Germany and may be contacted at andreas.jungherr@gmail.com.

Pascal Jürgens is a research fellow at the department of mass communication at the Johannes Gutenberg-Universität, Mainz, Germany and may be contacted at pascal.juergens@googlemail.com.

1. Introduction: online behavior and offline phenomena

The increasing use of the Internet and social media services has provided researchers with a new and rapidly growing data source on human behavior. Each interaction of users with online services (e.g. search engines, social networking sites etc.) leaves data traces, documenting online behavior. While most of these data traces remain inaccessible to researchers, some services (e.g. Google, Twitter, Facebook etc.) offer limited, yet structured access to some behavioral user data (e.g. through application programming interfaces, APIs). Other data, for example blog posts, or comments on commercial platforms, are publicly available and can be retrieved by automated scripts. These data document user behavior, user interactions, and allow inferences about user sentiments online at a scope and depth that in some respects, goes far beyond traditional data sources in the social sciences (Lazer et al., 2009). Although these data first and foremost document online behavior, they also hold the potential to illuminate offline phenomena (Rogers, 2009).

This realization has led to an impressive amount of research attempting to link data collected online to offline phenomena. Offline phenomena that have been linked to online data include economic indicators and sales figures (Gruhl et al., 2005; Choi and Varian, 2009), movie box-office results (Mishne and Glance, 2006; Asur and Huberman, 2010), the spread of diseases (Ginsberg et al., 2009), early-warning indicators of earthquakes (Sakaki et al., 2010), the play-by-play summary of sports and entertainment events (Chakrabarti and Punera, 2011; Shamma et al., 2011), the development of political protests (Jungherr and Jürgens, 2013), the development of political discourse (Weber et al., 2012), and even election results (Gayo-Avello, 2011). The data that was used for these analyses include the query logs of Internet search engines (Choi and Varian, 2009), posts and comments on blogs and commercial platforms (Balog et al., 2006; Mishne and de Rijke, 2006; De Choudhury et al., 2008), and messages on microblogging services (e.g. Twitter, Weibo et al.) (Shamma et al., 2011; Yu et al, 2011). Researchers often claim to be able to predict offline phenomena by analyzing patterns in the data provided by online tools (Choi and Varian, 2009; Asur and Huberman, 2010), or—more modestly—claim to be able to detect offline phenomena based on patterns in the data (Balog et al., 2006; Shamma et al., 2011; Weber et al., 2012). Others try to identify characteristic data patterns enabling the early detection of topics prone to generating a high volume of comments later on (Nikolov, 2012).

While most of these papers clearly show the potential that data produced by human interaction with online services hold for the analysis of offline phenomena, most of them are best understood as proofs of concept. This caveat also holds true for the empirical results presented further down. Still, in the first part of the paper we aim to move beyond this restriction by a systematic discussion of the relationship between online data and offline phenomena. We believe it is best to conceptualize online data traces as indicators of a user's interests at a given time. Thus, collective interests in topics or online content can be measured in the dynamics of time series based on the developments of variables documenting human behavior online (Kleinberg, 2003). The challenge is to connect these time series to offline phenomena.

The literature interested in the connection between online data and offline events has many subfields. The three approaches most relevant to our findings are concerned with correlation and prediction, event identification by activity spikes in time series, and the early detection of popular topics in online channels. The first approach examines if data on human behavior online systematically correlates with offline phenomena (e.g. Golder and Macy, 2011). Some authors even try to predict the occurrence of offline phenomena based on patterns in online data (e.g. Choi and Varian, 2009; Asur and Huberman, 2010). Some of these studies have gained significant attention, not least because of their original claims. Recently, however, the claims of some of the bolder studies—especially those with a focus on political outcomes—have met with severe criticism, since their results have been proven to be vulnerable to replication and scrutiny (Lui et al., 2011; Metaxas et al., 2011; Gayo-Avello, 2012; Jungherr et al. 2012).

Researchers interested merely in the identification of offline events with data on online behavior take a more modest route. Typically they use spikes in online data streams to identify offline events relevant to the online-service users under examination (Kleinberg 2002; Balog et al., 2006; Shamma et al., 2011; Weber et al., 2012; Jungherr and Jürgens, 2013). These bursts are detected by varying degrees of sophistication: for instance, the simple counting of frequent words represents one end of the spectrum (e.g. Twitter trending topics), while recording relative frequencies of wordstems during specific time intervals (e.g. Asur and Huberman, 2010; Shamma et al., 2011) represents another. The proof of concept we propose here adds another method for detecting of offline phenomena with online data.

As online tools become increasingly important so increases the interest of the public and traditional media in topics that are widely discussed online. Correspondently, the efforts to detect these “trending topics” (i.e. topics with temporally dense high-volume) early on intensify. Recently, an approach has been introduced (Nikolov, 2012) which predicts trending topics by means of non-parametric forecasting. In essence, volume curves of pre-established trending topics are clustered and compared to the volume development of new topics. The probability of a topic becoming a "trend" is estimated by its similarity to the development of past trending topics. This approach does not make any statements about the relation of online trends to offline phenomena. In fact, since it trains on past online phenomena, the method exclusively aims at mapping a set of specific patterns that relate to an algorithmically defined feature (trending). The method neither theorizes about external reasons for trend slopes (such as the speed of information diffusion, network structures etc.), nor does it relate to the entirety of topics constituting the communication space at any given time.

Researchers interested in using online data for the identification of offline events face obvious difficulties using Twitter data. First, every method focusing only on spikes in the total volume of Twitter messages, or only on the development of the most popular #keywords is bound to be diluted by #keywords without relevant information. As Table 1 shows, the most popular #keywords in a given time period are bound to contain their fair share of spam (e.g. #porn, #follow), Twitter usage conventions (e.g. #ff, #teamfollowback), and viral phenomena (e.g. #oomf, #500aday). Focusing only on the fluctuations of these #keywords would make any analysis very vulnerable to manipulations by interested actors who could infuse the Twitter communication sphere with a high volume of messages using specific #keywords (Metaxas and Mustafaraj, 2010; Metaxas and Mustafaraj, 2012). Thus, focusing only on fluctuations in the total amount of messages or the most popular #keywords holds limited appeal. Second, methods focusing on changes of word frequency in a given topical cluster of messages during a given period of time, therefore, promise to bring more precise results. Yet, these methods demand that researches choose relevant topics, word stems or hashtags beforehand. Thus, these methods promise to be precise, but are only helpful if researchers know beforehand what they are looking for.

[Table 1 about here]

The data pulse

We believe statistical models of a social media environment's "data pulse"—or normal state—are a valuable addition to the use of online data to identify offline events. We introduce the concept of the "data pulse" as the state of the system at a given time as determined only by known—and statistically modeled—aspects (e.g. the time of day, day of the week). This is an extension of a rudimentary idea already present in the literature (e.g. Balog et al., 2006; Hendrickson 2012). We show, that historical data of social media time series can be used to forecast the expected development of these time series. We do this not to predict the future, but to understand what future data patterns are to be expected, if tomorrow's users continue behaving as those of yesterday. If empirical data would significantly diverge from these forecasts, researchers would have an indicator that something out of the ordinary had happened in the time span in question. This might be an extraordinary event, which temporarily focused the attention of users, leading to a lower diversity of hashtags in use, when compared to forecasts based on user behavior in the past. Thus, we reverse the analytical approach for the prediction of future events. We aim less for the prediction of minute details in future developments. Rather, we intent to built forecasts based on basic, persistent patterns in human online behavior. To us, deviations between these forecasts and the empirical data are indicators of potentially relevant offline events or online phenomena. With this approach—identifying strays from the expected "data pulse" of a social media system and determining their nature afterward—we propose a method that promises to be reasonably stable against spam and to allow researchers the detection of unexpected topics potentially relevant for the detection offline phenomena in online data.

In this paper, we suggest to use data, documenting user behavior online, to determine the "normal state" of a social media information environment—the pulse of a data source—to forecast a normal state of the system at a different time. If the empirical data deviates from this forecast too strongly, we take this as an indicator that the activities of users on a given channel in the time span in question deviate from their normal behavior. Thus, our forecast of the normal state serves as a benchmark against which to hold empirical data. If the gap between forecast and empirical data becomes too big, we may determine whether something out of the ordinary has happened that captured the attention of the users of a given channel. It is important to note that the events potentially captured by this approach are not unexpected events, in the sense of "black swans"—completely unexpected events with potentially strong impact (Taleb, 2007). Instead, these events are special in that they capture unexpectedly high levels of attention by Twitter users. This could be unexpected events (e.g. the London riots) or scheduled, mediated events (e.g. coronations, sports, or debates in Presidential races). These media events (Dayan and Katz, 1992) are clearly not unexpected. In fact, they are scheduled and widely promoted well in advance. They are unexpected in the sense that they lead users to deviate from their established communication patterns and make them focus their attention on these events. Thus, these events break the "data pulse" of expected user behavior.

We start with the discussion of various types of online services used by researchers to analyze offline phenomena and the various data generating processes connected to them. We then discuss types of offline phenomena that in the past have been linked to specific patterns in online data. Then we introduce the concept of a "data pulse" and our reasons for using data collected on the microblogging platform Twitter. In a final step, we build a proof of concept to show the potentials that arise by forecasting an expected "data pulse" and by comparing it with empirical data. Large differences between the forecasted normal state and the empirically measured data serve to identify extraordinary events offline and online that were relevant to users of the respective online service.

2. Data on online behavior and connections to offline phenomena

A steadily increasing number of people use social media tools habitually in everyday life. This activity produces an ever-increasing amount of data that promise social scientists insights into human behavior and human interaction in increasing depth and scale. But these data do not only document human behavior. They also offer insight into what topics, people, or events held the attention of social media users at a given moment. Thus, Information and Communication Technologies (ICT) become sensors, documenting the varying objects of attention of their users. Still, the question remains what aspect of human behavior these data document.

The Pyramid of Involvement

As the interests of ICT users are connected to, or were sparked by, offline phenomena, ICT data traces might hold information on these offline phenomena. As different online services require varying levels of engagement by the user, it is sensible to assume that data collected from different online services document different behaviors by different subgroups of the population. ICT can diagrammed as a pyramid of involvement. In Figure 1 we have depicted three particularly important ICT on this pyramid.

[Figure 1 about here]

Search engines probably hold the lowest requirements of prior engagement by their users. To conduct her interaction with the service, a user interested in a given topic, must only enter a search term appropriate to her immediate interests. This should lead to a wide variety of interests being documented in the logs of online search queries. Another characteristic of search engines is that they are widely used. Thus, the data traces of the interactions by their users might document interests of a wider part of the population than data collected by online services with a narrower user base. Although search engines are used by 59% of all American adults on a daily basis, the user demographic is highly skewed towards "younger, more educated and more affluent search engine users" (Purcell et al., 2012). While search engines show the highest daily usage of online tools by Americans, those using these services frequently are part of a very specific demographic. Their interests—mirrored by the data traces of their online use—should thus not be considered representative of the interests of *all* Americans.

In general terms, publishing a status message on a social networking site, like Facebook, or a microblogging service, such as Twitter or Weibo, suggests a higher level of involvement than performing a simple search. Publishing a message for a private or semi-public audience furthermore suggests that a user actively wants to share her thoughts on a topic or her reactions to an event. Therefore, data collected on these platforms document a different aspect of human behavior than search-query logs. Consequently, it is important to recognize that social-networking sites and microblogs are used by a considerably smaller and demographically more specific user base than online search engines (Correa et al., 2010; Mislove et al., 2011; Smith and Brenner, 2012). The results of analyses based on these data might speak for merely a fragment of the population, namely younger, more educated and affluent people.

In contrast, a comparatively higher level of involvement is necessary for a user to blog about a topic or an event. This is also true for comments or product reviews on commercial platforms, such as amazon.com. This hypothesis is supported by the fact that the percentage of internet users who actively blog or post comments is smaller than the percentage of those who post status messages on microblogs or social networking sites, which turn, is smaller than the percentage of internet users using search engines (Busemann and Gscheidle, 2011). Demonstrably, analyses based on data collected on blogs or comment forums capture the behavior and interests of potentially highly involved persons that merely form a small, non-representative subgroup within the larger population.

These general observations correspond with surveys of user behavior online. For example, the Pew Internet & American Life Project found that on a typical day 59% of Americans use a search engine to find information,[1] 8% use the microblogging service Twitter,[2] 4% create or work on their own online journal or blog[3] and 4% post a comment or review online about a product they bought or a service they received[4] (Pew 2012). Since these statistics come from different surveys in different years, they can only serve as rough indicators for usage patterns. Still, similar patterns can be found in other countries (e.g. Busemann and Gscheidle, 2011). To us, these usage patterns indicate the aforementioned "pyramid of involvement".

The differences documented by the pyramid of involvement have to be consciously addressed when analyzing data generated by the interaction of humans with online services. On the most basic level, the pyramid of involvement suggests that results based on data from different services will probably *not* be interchangeable. It is also important to remember that these data can only speak for the behavior and interests of the users of the online service that the data was collected on. Thus, the inference of behavior or interest of the whole population is not necessarily possible.

Data generating processes

[1] Survey date: February 2012.

[2] Survey date: February 2012.

[3] Survey date: May 2011.

[4] Survey date: September 2010.

Data collected on different ICT are not only different based on the level of involvement motivating their users to use the services. The data also differ according to the discrepancies between the various data-generating processes. These deviations have consequences for the type of offline phenomena that might be traced or analyzed using said data.

On the one hand, we have data from devices and services that document the behaviors of users who—probably—unintentionally left data traces. One example for this type of data might be a log of geographic locations of mobile-phone users. Another example might be a search-engine log that documents the queries users have entered. Usually these data also contain metadata (e.g. the device on which the search was performed, the geographic location of the search etc.). Principally, these data can be used to determine the interests and habits of users at pre-set locations at given times (e.g. a political election, or electricity prices). Naturally, user intent is very closely linked to the available data on user behavior. To give an example, we may assume that if users want to get information on a candidate in an election, they might use the candidate's name, the name of the party, or the campaign slogan as search terms for online searches. The use of such parameters occurs irrespective of user's attitudes toward the candidate or the candidate's ideological positions. Thus, the logs of this search engine document the user's interest in this candidate, but do not necessarily document her attitude towards the candidate (see Figure 2). Users create these data passively in the process of using an online service. These accidental, passively created data have been used in the past to predict mobility patterns in metropolitan areas (González et al., 2008), track the spread of diseases (Ginsberg et al., 2009), predict market prices (Choi and Varian, 2009) and identify the development of public discourse on politics (Weber et al., 2012).

[Figure 2 about here]

There are also data that are intentionally published by users, for example status messages on profiles on social networking sites or microblogs. These data document the full text of the status messages, links to content on the Internet, and links to other users mentioned in these messages. These messages also contain metadata (e.g. the device on which the messages were written, the geographic location the messages were written at etc.). In principle, these data can be used to determine those interests and personal connections that users want to document publicly at a given place and a given time. This is a crucial difference between these data and those created passively by users. To return to the example above, we may assume that a user might be interested in an upcoming election and so she uses her microblogging account to voice her support for a candidate. Her message might identify the candidate, the party or a campaign slogan. These data might resemble the data collected in the logs of her online searches; however, they are created by different user motivations. For a user to intentionally publish a message about a topic, the user must have a strong motivation. In the case of political candidates, this motivation usually consists of either support or opposition. Accordingly, these data are highly dependent on the motivation of a user to see them published. Intentionally published data have been used to document the spread of earthquakes (Sakaki et al., 2010), coordinate disaster relief efforts (Verma et al., 2011), and to identify the sequence of micro-steps in mediated events, and political protests (Shamma et al., 2011; Jungherr and Jürgens, 2013). Distinguishing between the different data-generating processes of individual online services is important because these distinctions determine what inferences researchers can validly draw on offline phenomena based on online data.

[Figure 3 about here]

Offline Phenomena

There are different types of offline phenomena that have been connected to online data. These differences are based on the link between a phenomenon and an online data source. Some offline phenomena are directly linked to the data found on online channels. For example, if an Internet user posts a status message on her microblogging profile, "I am sick with the flu, now going out to buy medicine", this message is directly linked to her being sick with the flu and thus a valid indicator for everyone analyzing the spread of influenza through the relative frequency of flu related keywords in online data. Other examples for offline phenomena directly linked to online data are online reactions (e.g. status messages, search queries et al.) to events, and online data documenting users' interests in buying or selling goods. To put this point more generally, every offline phenomenon that directly leads a user to interact with an online service can be understood as directly linked to the resulting online data. In turn these data can be used for the identification and—depending on the quality of the metadata—temporal and geographic tracking of the underlying offline phenomena.

Other offline phenomena are indirectly linked to online data. Some examples are the attempt to determine the collective mood of a population to predict the development of stock prices (Bollen et al., 2011), or commercial success of films or goods based on the level of topically related word-of-mouth chatter (Asur and Huberman, 2010). In the first example, researchers assume that the emotional mood of a population is somehow linked to the economic state (for example that collective mood swings are to be understood as premonitions of the collective unconscious of coming economic ups and downs). In this case, the collective mood might show in the phrasing of content published online. Some researchers claim that they can measure the dynamics of collective mood and use them to predict upcoming movements of economic indicators (Bollen et al., 2011). If this relationship proved to be robust, one could characterize this relationship between offline phenomena and online data as indirect.

There are also offline phenomena that have no relationship with online data, but that have, nevertheless, been object of analyses or predictions based on online data. This is very obvious in the prediction of election results based on varying measures of the intensity of political conversations online (Tumasjan et al., 2010). A naïve observer could assume that the connection between topical mentions of names of political parties or candidates could be seen as an indirect link, comparable to the one described above. Yet closer inspection shows this to be a misconception. Elections are decided by the votes of vocal and non-vocal supporters of parties and politicians alike. An approach just focusing on the opinions and statements of vocal political supporters ignores the political convictions of those silent in the public discourse but willing to vote. As it turns out, the convictions between the politically non-vocal voters sometimes deviate from the opinions of those voicing their political opinions before an election, and thus the parties supported by non-vocal supporters might be underestimated. In communication research, this phenomenon has become known as the “spiral of silence” (Noelle-Neumann, 1991). Recently similar patterns have been shown for political discourse online (Mustafaraj et al., 2011). Thus, the level of chatter published online or offline stands not necessarily in a meaningful relationship with the outcome of political elections. Accordingly the practice of predicting election results based on data collected online has—after a short series of highly publicized papers—become increasingly discredited (Gayo-Avello, 2012).

3. Forecasting the data pulse

After this general discussion, we present an event-detection approach, designed to be resilient to spam, for analyzing social media data. This approach should allow for the detection of topics and events that created strong reactions online without researchers having to know relevant hashtags beforehand. The state of a social media channel at any given time can be quantified by measuring a set of variables. This could be the total amount of messages on a microblogging service posted during a given time span, the number of users actively posting messages, the relative frequencies of specific terms or—as in the following proof of concept—the diversity of hashtags. These variables and their historic states can be documented in time series, which in turn can serve as source for statistical analyses and forecasts. This is important, because time-series analysis allows us to identify regular patterns in the data (e.g. seasonal trends in the data) and to model a normal state of the system. By "normal state"—or "data pulse"—we mean the state of the system at a given time (e.g. diversity of hashtags during a given time span) as determined only by known—and statistically modeled—aspects (e.g. the time of day, day of the week). If we compare this forecast with empirical data measuring the state of the variable in question at the time of interest we are able to identify deviations in the trend—the difference between the value as forecasted sans seasonality and the value as measured. Depending on the levels of these deviations, we are then able to determine if, at that time in question, the actual state of the system significantly deviated from the expected, or normal, state. If this were the case, we would expect that the reason for this deviation would either lie in our incomplete knowledge about the dynamics of the time series, or that online or offline phenomena have led users to deviate from their usual usage patterns. Deviations from the "data pulse" could thus be used as indicators of the occurrence of phenomena—originating online or offline—relevant to the users of the service in question.

We show the potential of this approach by using data collected on the microblogging service Twitter, but in principle our approach should also work for different data sources. Twitter is an online service that allows users to post short text messages of up to 140 characters in length on personalized profiles. Users can also "follow" other accounts, which means that they can subscribe to new messages of other users and can access new messages by all users followed by them in an aggregated message feed. We choose Twitter for our analysis since the service has a relatively open data-access policy, enabling us to use a considerable volume of data for this analysis. Besides the ease of access, Twitter data has characteristics that make it a promising source for the analysis of offline phenomena. Twitter started out as a medium for self-expression but has since evolved into a forum for discussing current events and politics. It has therefore become a promising data source for researchers interested in these topics (Kwak et al., 2010; An et al., 2011).

Another characteristic of Twitter facilitates the automated analysis of its data: since Twitter messages are restricted in length to 140 characters or less, users developed cultural practices that helped to establish the context of a tweet with only a few characters. Some of these practices are documented in Table 2. It has been shown that there are regularities in the use of these usage practices (Huberman et al., 2009; Cha et al., 2010). Thus, it seems reasonable to assume that the relative frequency of the occurrences of these cultural practices offers a promising base to model the "data pulse" of Twitter use.

[Table 2 about here]

In our proof of concept, we decided to focus fluctuations in the diversity among the 1.000 most popular hashtags during February 2012 and June 2012.[5] In doing so, we avoid some pitfalls inherent in approaches focusing on other examples, such as the fluctuation of the volume of messages, or the volume of selected hashtags. As with other online channels, spam constitutes a considerable component of the data found on Twitter. A surprising amount of messages can be directly linked to spam and/or marketing activity (see for example Table 1). This can potentially obstruct event-detection with social media data. What is more, spam volumes can rise and fall drastically, and thus pose a challenges to methods that use high-volume levels and drastic volume changes to identify meaningful time spans. In order to identify

[5] We chose to limit the number of tags mainly because their volume distribution exhibits a very long tail. Hence, there are many hashtags that do not contribute significantly to the overall communication but, by vastly extending the number of available categories, severely dampen the calculated diversity index.

meaningful developments in the Twitter stream, topic-bound methods need to reliably identify and filter out spam.

Alternatively, one could focus just on the development of one hashtag relevant to the topic under examination. But there is a further problem with models that only include the time series of one topic: they will always fall short when (latent) events are reflected in several different topics. This shortcoming can be seen in one of our later examples where the hashtags "#superbowl" and "#giants" are both caused by the US Super Bowl event. In fact, it is easy to imagine topic constellations in which several smaller, thematically connected, yet fragmented, topics are not identified by algorithms because they were "overshadowed" by a single dominant, homogenous topic. In further studies, it might be possible to use the degree of topical fragmentation to separate potentially meaningful from less important topics. What if superfluous fads (such as online memes) would exhibit a high topical homogeneity, whereas widespread political events gave rise to a plethora of connected but different hashtags?

To avoid these issues, we focus on the fluctuations in the diversity of hashtags used during a given time interval. By measuring the relative volume of messages containing each hashtag and comparing it with the volume of all other active hashtags, we can quantify the spread of attention across topics. This spread of prevalence is usually called “diversity.” This concept has been used across various disciplines such as ecology, finance, mass communication, and others (McDonald & Dimmick, 2003). For our purposes, one particular dynamic of the diversity-time series seems especially meaningful. If the spread of attention shifts, that is, if the percentage share of one or several topics grows at the cost the share of others, then the total diversity value decreases. Following our assumptions, we argue that this will happen only when external forces (e.g. events, memes) lead users to change their communicative behavior. In essence, drops in diversity represent phases during which Twitter users refrain from using many of the possible hashtags in favor of mentioning one or more dominant ones. An increase in diversity, on the other hand, could be most convincingly attributed to a sudden influx of new users who mention a set of hashtags that was not present before, thereby spreading activity across categories (hashtags).

The diversity of Twitter topics was measured using a dataset of 3.788.651.747 messages gathered during the months of February through June 2012, using an extended access level of Twitter's random sample stream. This was formerly specified to include 10% of all tweets but subsequently reduced; in any case the data averages about 18.000 messages per minute.[6] We extracted the amount of messages per hashtag per hour. Since a large proportion of tags was used very infrequently and thus had both a low impact and dampening effect on diversity, we selected the 1.000 tags with the highest volume. The distribution of messages per hashtag displays a long tail. While there are very few frequently used tags, most are used very infrequently. Calculating a diversity index over all tags would—due to the large number of categories—yield scarcely any fluctuation. We therefore selected only the top 1.000 hashtags for our analysis. The total range of messages per tag goes from one for tags used only once to 5.648.780 for the tag used most often. Our sample of the top 1.000 tags represents hashtags from 53.821 occurrences upwards, and thus spans over 99% of the total range. Diversity was calculated on this dataset using Shannon's diverstiy index, sometimes called SDI or "Shannon's H" (Shannon 1948, McDonald & Dimmick, 2003):

$$H' = -\sum_{i=1}^{R} p_i \, ln \, p_i$$

Since diversity is expressed at any given time by a single value, this value produces an univariate time series and can be examined with established methods for time-series analysis. In order to keep the approach as simple as possible, we opted to use STL, a seasonal trend decomposition technique based on Loess (Cleveland et al., 1990). Using the "stl" package in R (R Development Core Team, 2011), the diversity time series were decomposed into a seasonal component, a trend, and a remainder. Major drops in diversity can thus be identified visually in the trend component of the resulting graphs.

In this proof of concept, we used the by far most basic version of this approach. For extensions of this approach, it is easily imaginable to automate the identification of significant diversity decreases (e.g. through a parametric procedure).

[6] The license granted by Twitter Inc. for the use of its data explicitly prohibits the sharing of raw data, even in parts. Still, to enable the replication of our approach, we produced two csv files, documenting the total number of messages per day and the daily usage counts of top 1.000 hashtags. These files are accessible at https://github.com/trifle/twitter-diversity.

Also, instead of calculating the diversity of the entire Twitter stream, other analyses (e.g. on sub-topical, local or community-based levels) are easily imaginable.

To facilitate the visual inspection, we will take two months from the total time-span and examine some cases of sudden drops in diversity in greater detail. The figure for February (Figure 4) shows the three elements of the STL on a shared time scale. The top panel shows the original time series, labeled “data”. Below, the “seasonal” component represents the extracted pattern of daily variation, which is common to all cycles in the dataset. The third panel displays a smoothed trend component over time. Finally, the “remainder” represents residuals that were not captured by either seasonal component or smoothed trend. Given that diversity is bound to drop when the composition of topics changes, we expect the trend to display significant drops at points in time when a major event displaces other topics in communication.

[Figure 4 about here]

Examining the trend of the data at hand, we see various deviations. Some of them may be the result of yet-undetected seasonalities, technological constraints, and many should be the result of the stochastic nature of measurements of human behavior. Yet, there are several large deviations from the null horizon. Both edges of the time series fluctuate somewhat. While the overall volume may drop due to technical errors (either with Twitter’s service or the computer used to capture the data) that may lead to brief periods of missed tweets, the diversity measure is relatively stable, as long as the errors omit messages uniformly and do not drop to zero (as is the case in the middle of February). Beyond these features, many smaller increases stay unexplained. Our basic premise is that modeling the regular features of the diversity of message topics reveals relevant events beyond our ability to forecast. This holds true if we find that drops in the trend timeline actually correspond to explainable, singular phenomena. Looking at the trend series, there are several notable drops, for example around the 6th and 8th. Upon investigation of the topic compositions, we find that the 6th is marked by a spike in a TV-related hashtags (the voice, "#voice"), most prominently by the Super Bowl ("#superbowl" and "#giants"). As seen in the stacked plot (Figure 5) that highlights the five most prominent tags during this time span, the Super Bowl topics expand upwards and downwards which indicates a displacement of other topics. The result here is rather sobering, given that February was a month that showed intense conflict in the Syrian civil war. If political events were met with at least some interest from the majority of Twitter users, we would have expected to see this reflected in the diversity. Instead, the topic seems to have lost out against topics with stronger relevance to western Twitter users.

[Figure 5 about here]

Among the other deviations, we find the 2012 UEFA European Football Championship or “euro2012” in short. Corresponding to the days of prominent matches (Figure 6), we find that this topic dominates the topic composition. The effect of displacement of other topics is quite clear on the 24th of June:

[Figure 6 about here]
[Figure 7 about here]

Given that in these examples the method successfully detected actual events, we furthermore wanted to test whether it would be able to identify other meaningful developments. In this respect, the limits of the diversity metric become quite apparent: although the 16th and 17th of June saw important elections in Egypt (a country that was prominently discussed on Twitter before), those do not register in the trend. Looking at the volume, the hashtag #egypt never reaches beyond 4.000 Tweets per hour, easily being drowned by tweets referencing the European Soccer Championship.

By design, diversity is a rather coarse indicator in that it detects a contraction of attention. Insofar, its trends identify situations in which the twittersphere becomes increasingly focused on one singular topic. In order to better quantify the actual impact of each topic, so as to get a more accurate picture, not only of the singular dominant topic, but also of smaller ones, a simple technique could be proposed: one merely needs to perform a calculation of the diversity, with each topic removed in turn. Topics can then be ranked by the amount by which their removal increases the metric.

To conclude, we showed how a series of very simple analytical steps can model meaningful causes for variations in the structure of hashtags actively used on

Twitter. In doing so, we were able to explore, explain, and formalize sensible assumptions about behavioral patterns in the data documenting digital traces of user behavior online. By using the STL technique, we gained a time series with drastically reduced fluctuations, which represents deviations from our expectations. Under this approach, it is easily possible to add ever more model steps onto each other, each in turn reducing the variance of the remainder as it explains a bit of “what’s going on” and thus to come to an increasingly realistic model of the “data pulse”.

With this proof of concept we add to the growing literature on event detection with social media data. In our two very simple examples we showed that a diversity based approach to event detection was successful in detection attention shifts in Twitter data towards the Super Bowl and the football Eurocup 2012. One could argue that these events were big sporting events that did not exactly need shifts in Twitter users’ behavior to publicize their occurence. Still, this event-detection method could easily be used on smaller populations, for example political activist or journalists, to detect changes in their attention manifest in their Twitter activity. The applications for social scientists are obvious: the approach enables the identification and tracking of topics that gained attention by a given population of Twitter users going well beyond what could be accounted for by their behavior in the past. This allows researchers to account for changes in the user behavior of populations relevant for social scientists and the identification of topics that were of unusual interest to them during a given time span.

4. Conclusion

In this paper we have shown that data produced by human interaction online holds strong potential for social scientists interested in offline phenomena. As shown above many empirical studies have illustrated the potential of these data in specific case studies. We think that now is the time to move beyond isolated, individual cases and think more systematically about the theoretical links between data collected on different online services and various offline phenomena.

In the first part of this paper we added to the discussion about links between online data and offline phenomena. We did so by suggesting three elements of online services, which have to be taken into account in the interpretation of analyses based on data collected on these services. These elements are the pyramid of involvement, the different data generating processes of each service, and the characteristic of the link between the human interaction with the online service and the offline phenomena in question.

In the second part of the paper we showed the potential to detect extraordinary phenomena by investigating externally induced changes in the “normal state” of online data based on topic diversity. With this approach, we contribute to the discussion of the potential of forecasts based on data collected online. But instead of attempting to forecast exceptional phenomena—for example by identifying typical patterns in the data—we suggest to model the “normal state” of a system and take strong differences between the model and the empirical data as an indicator for the occurrence of extraordinary offline phenomena that led users to change their behavior mirrored by the unexpected patterns in the empirical data.

5. References

An, J., Cha, M., Gummadi, K. and Crowcroft, J. (2011), “Media landscape in Twitter: A world of new conventions and political diversity”, in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM) in Barcelona, Spain, July 2011*, The AAAI Press, Menlo Park, California.

Asur, S. and Huberman, B. (2010), “Predicting the Future with Social Media”, in *WI-IAT '10 Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, IEEE, Washington, DC, pp. 492-499.

Balog, K., Mishne, G. and de Rijke, M. (2006), “Why are they Excited? Identifying and Explaining Spikes in Blog Mood Levels”, in *11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL 2006), April 2006*, Trento, Italy.

Bollen, J., Mao, H. and Zeng, X. (2011), “Twitter mood predicts the stock market”, *Journal of Computational Science*, Vol. 2 No. 1, pp. 1-8.

Busemann, K. and Gscheidle, C. (2011), “Web 2.0: Aktive Mitwirkung verbleibt auf niedrigem Niveau”, *Media Perspektiven*, No. 7-8, pp. 360-369.

Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, K. (2010), “Measuring User Influence in Twitter: The Million Follower Fallacy”, in *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM), May 2010*, The AAAI Press, Menlo Park, California.

Chakrabarti, D. and Punera, K. (2011), "Event summarization using Tweets", in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM) in Barcelona, Spain, July 2011*, The AAAI Press, Menlo Park, California, pp. 66-73.

Choi, H. and Varian, H. (2009), "Predicting the present with Google Trends", working paper, Google Inc., Mountain View, California, available at http://ec.europa.eu/bepa/pdf/seminars/google_predicting_the_present.pdf (accessed 16 June 2012).

Cleveland, R. B., Cleveland, W. S., McRae, J. E. and Terpenning, I. (1990), "STL: A seasonal-trend decomposition procedure based on loess", *Journal of Official Statistics*, Vol. 6 No. 1, pp. 3–73.

Correa, T., Hinsley, A. W. and de Zúñiga, H. G. (2010), "Who interacts on the Web?: The intersection of users' personality and social media use", *Computers in Human Behavior*, Vol. 26 No. 2, pp. 247-253.

Dayan, D. and Katz, E. (1992), *Media Events: The Live Broadcasting of History*. Harvard University Press, Cambridge, MA.

De Choudhury, M., Sundaram, H., John, A. and Seligmann, D. D. (2008), "Can blog communication dynamics be correlated with stock market activity?", in *HT '08 Proceedings of the 19th ACM Conference on Hypertext and Hypermedia*, ACM, New York, NY, pp. 55-60.

Gayo-Avello, D. (2011), "Don't Turn Social Media Into Another 'Literary Digest' Poll", *Communications of the ACM*, Vol. 54 No. 10, pp. 121-128

Gayo-Avello, D. (2012), "'I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper': A Balanced Survey on Election Prediction using Twitter Data", available at http://arxiv.org/pdf/1204.6441v1.pdf (accessed 16 June 2012).

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. and Brilliant, L. (2009), "Detecting influenza epidemics using search engine query data", *Nature*, Vol. 457, pp. 1012-1014.

Golder, S. A. and Macy, M. W. (2011), "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures", *Science*, Vol. 333 No. 6051, pp. 1878-1881.

González, M. C., Hidalgo, C. A. and Barabási, A. L. (2008), "Understanding individual human mobility patterns", *Nature*, Vol. 453, pp. 779-782.

Gruhl, D., Guha, R., Kumar, R., Novak, J. and Tomkins, A. (2005), “The Predictive Power of Online Chatter”, in *KDD '05 Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, ACM, New York, NY, pp. 78-87.

Hendrickson, S. (2012), “Social media pulse: the shape of breaking news on social media”, working paper, Gnip, Inc., Boulder, Co, 12 July.

Huberman, B. A., Romero, D. M. and Wu F. (2009), “Social networks that matter: Twitter under the microscope”, *First Monday*, Vol. 14 No. 1, available at http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2317/2063 (accessed 16 June 2012).

Jungherr, A. and Jürgens, P. (2013), “Stuttgart’s black Thursday on Twitter: Mapping Political Protests with Social Media Data”, in Gibson, R., Cantijoch, M. and Ward, S. (Ed.), *Analyzing Social Media Data and Web Networks: New Methods for Political Science*, Palgrave Macmillan, New York, NY.

Jungherr, A., Jürgens, P. and Schoen, H. (2012), “Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. ‘Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment’”, *Social Science Computer Review*, Vol. 30 No.2, pp. 229-234.

Kleinberg, J. (2003). “Bursty and hierachical structure in streams”, *Data Mining and Knowledge Discovery*, Vol. 7 No. 4, pp. 373-397.

Kwak, H., Lee, C., Park, H. and Moon, S. (2010), “What is Twitter, a social network or a news media?”, in *WWW '10 Proceedings of the 19th International Conference on World Wide Web*, ACM, New York, NY, pp. 591-600.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. and Van Alstyne, M. (2009), “Computational social science”, *Science*, Vol. 323 No. 5915, pp. 721-723.

Lui, C., Metaxas, P. T. and Mustafaraj, E. (2011), “On the predictability of the U.S. elections through search volume activity”, paper presented at the e-Society Conference, March, 2011, Avila, Spain, available at http://cs.wellesley.edu/~pmetaxas/e-Society-2011-GTrends-Predictions.pdf (accessed 16 June 2012).

McDonald, D. G. and Dimmick, J. (2003), “The conceptualization and measurement of diversity”, *Communication Research*, Vol. 30 No. 1, pp. 60–79.

Metaxas, P. T. and Mustafaraj, E. (2010). “From obscurity to prominence in minutes: political speech and real-time search”, in *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, April 26-27th, 2010, Raleigh, NC: US.

Metaxas, P. T. and Mustafaraj, E. (2012). “Social media and the elections”, *Science*, 338, pp. 472-473.

Metaxas, P. T., Mustafaraj, E. and Gayo-Avello, D. (2011), “How (not) to predict elections”, in *Proceedings of the 2011 IEEE 3rd International Conference on Social Computing*, IEEE, Washington, DC, pp. 165-171.

Mishne, G. and de Rijke, M. (2006), “Capturing Global Mood Levels Using Blog Post”, in *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, The AAAI Press, Menlo Park, California.

Mishne, G. and Glance, N. (2006), “Predicting Movie Sales from Blogger Sentiment”, in *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, The AAAI Press, Menlo Park, California.

Mislove, A., Lehmann, S., Ahn, Y. Y., Onnela, J. P. and Rosenquist, J. N. (2011), “Understanding the demographics of Twitter users”, in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM) in Barcelona, Spain, July 2011*, The AAAI Press, Menlo Park, California.

Mustafaraj, E., Finn, S., Whitlock, C. and Metaxas, P. T. (2011), “Vocal minority versus silent majority: discovering the opinions of the long tail”, in *Proceedings of the 2011 IEEE 3rd International Conference on Social Computing*, IEEE, Washington, DC.

Nikolov, S. (2012), *Trend or no trend: a novel nonparametric method for classifying time series (Master’s thesis)*. Massachusetts Institute of Technology, Cambridge, MA.

Noelle-Neumann, E. (1991), “The theory of public opinion: the concept of the spiral of silence”, in Anderson, J. A. (Ed.), *Communication Yearbook 14*, Sage, Newbury Park, CA, pp. 256-287.

Pew Internet & American Life Project (2012), “What Internet Users Do On A Typical Day”, available at http://pewinternet.org/Trend-Data-(Adults)/Online-Activities-Daily.aspx (accessed 29 September 2012).

Purcell, K., Brenner, J. and Rainie, L. (2012), “Search Engine Use 2012”, Pew Internet & American Life Project, available at http://pewinternet.org/~/media//Files/Reports/2012/PIP_Search_Engine_Use_2012.pdf (accessed 16 June 2012).

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, available at http://www.r-project.org/ (accessed 16 June 2012).

Rogers, R. (2000), *The End of the Virtual: Digital Methods*, Amsterdam University Press, Amsterdam, Netherlands.

Sakaki, T., Okazaki, M. and Matsuo, Y. (2010), “Earthquake shakes Twitter users: real-time event detection by social scensors”, in *Proceedings of the 19th international world wide web conference, WWW '10 Proceedings of the 19th International Conference on World Wide Web*, ACM, New York, NY, pp. 851-860.

Shamma, D. A., Kennedy, L. and Churchill, E. F. (2011), “Peaks and persistence: Modeling the shape of microblog conversations”, in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, ACM, New York, NY, pp. 355-358.

Shannon, C.E. (1948). “A Mathematical Theory of Communication”, *The Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656.

Smith, A. and Brenner, J. (2012), “Twitter Use 2012”, Pew Internet & American Life Project, available at http://pewinternet.org/~/media//Files/Reports/2012/PIP_Twitter_Use_2012.pdf (accessed 16 June 2012).

Taleb, N. N. (2007), *The Black Swan: The Impact of the Highly Improbable*. Random House, New York, NY.

Tumasjan, A., Sprenger, T. O., Sander, P. G. and Welpe, I. M. (2010), “Predicting elections with Twitter: what 140 characters reveal about political sentiment”, in *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM), May 2010*, The AAAI Press, Menlo Park, California, pp. 178-185.

Verma, S., Vieweg, S., Corvey, W. J., Palen, L, Martin, J. H., Palmer, M., Schram, A. and Anderson, K. M. (2011), “Natural language processing to the rescue? Extracting ‘Situational Awareness’ Tweets during mass emergency”, in

Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM) in Barcelona, Spain, July 2011, The AAAI Press, Menlo Park, California, pp. 386-392.

Weber, I., Garimella, V. R. K. and Borra, E. (2012), "Mining Web Query Logs to Analyze Political Issues", in *Proceedings of ACM WebSci'12*. ACM, New York, NY.

Yu, L., Asur, S. and Huberman, B. A. (2011), "What Trends in Chinese Social Media" paper presented at the 5th SNA-KDD Workshop'11 (SNA-KDD'11), August 21, 2011, San Diego CA USA, available at http://arxiv.org/pdf/1107.3522v1.pdf (accessed 16 June 2012).

Tables:

Table 1: The twenty #keywords with the highest volume between February and June 2012

#keyword	**Number of uses**
#ff	5.648.780
#teamfollowback	5.304.230
#np	4.078.016
#oomf	3.875.364
#rt	3.863.284
#nowplaying	2.306.588
#nf	1.705.962
#bahrain	1.683.564
#followme	1.681.314
#followback	1.586.823
#fb	1.552.503
#1	1.377.774
#porn	1.373.942
#follow	1.335.680
#500aday	1.303.061
#sougofollow	1.229.433
#followmejp	1.117.858
#tfb	1.087.211
#bakugeki	1.086.015
#retweet	1.079.811

Table 2: Elements of a Twitter message

Data contained in a Tweet	Description
Text of message with relevant topical keyterms and modifiers	
username	every Twitter user has an unique username, that is referenced in each message
date	the metadata of each Twitter message contains the exact time, date and time zone when the message was posted
location	the metadata of each Twitter message contains the location where the message was written and posted (if enabled by the user)
@message to another user	to identify messages that are part of a public conversation between two Twitter users, Twitterers precede the text of their message with the username of the addressee preceded by an "@" (i.e. @username)
@mention of another user	if a user is not directly addressed but mentioned in a tweet the @username convention is used in the text of the message instead of the beginning
RT verbatim	the retweet (RT) is a convention in which Twitter users copy messages of other users verbatim and precede these messages by the character string "RT @username"
RT modified	it is also possible to retweet a message and commenting on it in the same tweet
#keywords	users can mark their messages with keywords proceeded by the "#" sign, this is often done to explicitly anchor a tweet in a specific topical context
links to other web content (e.g. websites, pictures, videos et al.)	messages can contain (often shortened) links to other content on the web

Figures:

Figure 1: The pyramid of involvement of different online services

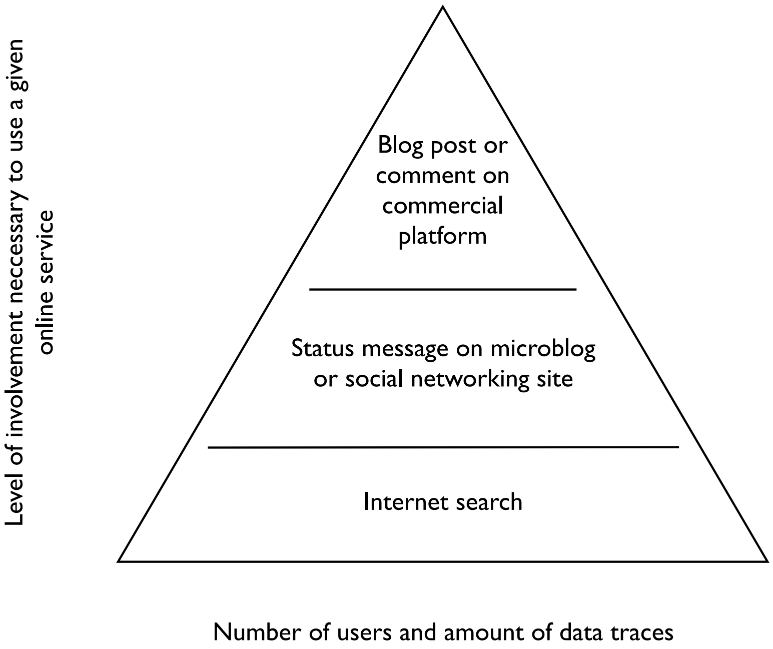

Level of involvement neccessary to use a given online service
Blog post or comment on commercial platform
Status message on microblog or social networking site
Internet search
Number of users and amount of data traces

Figure 2: Data generating process of search engine logs

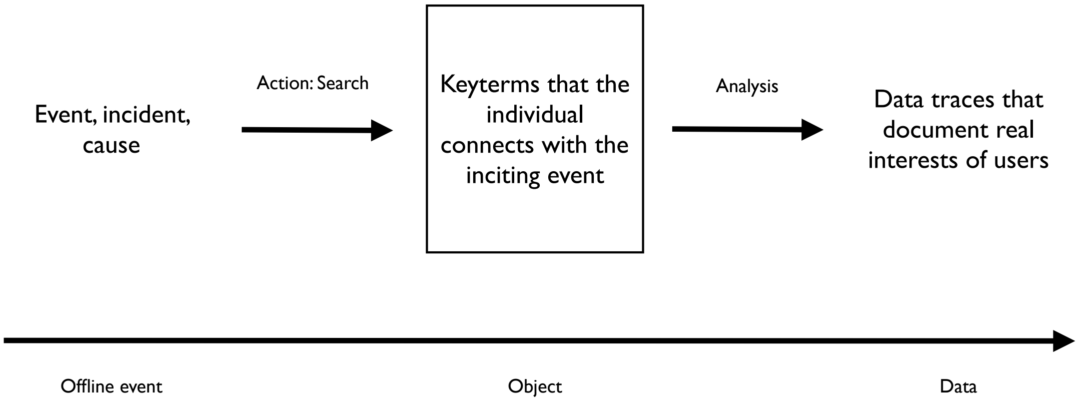

Event, incident, cause
Action: Search
Keyterms that the individual connects with the inciting event
Analysis
Data traces that document real interests of users
Offline event
Object
Data

Figure 3: Data generating process of microblogging data

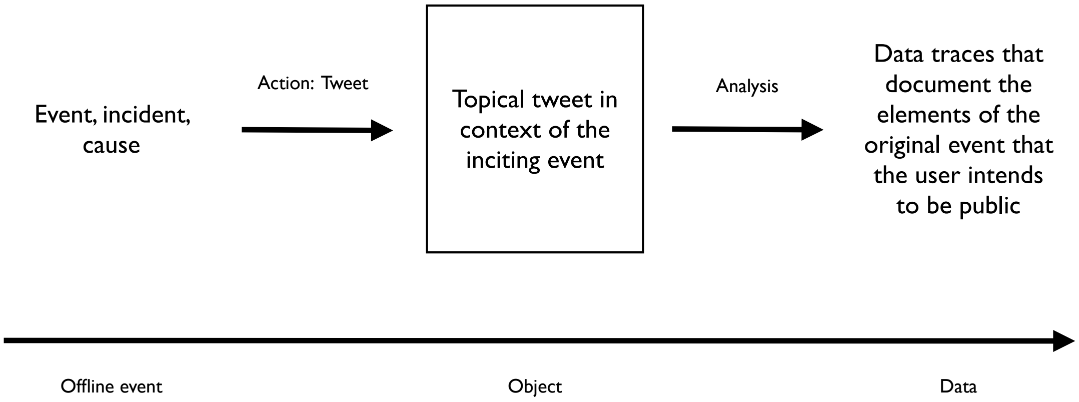

Event, incident, cause
Action: Tweet
Topical tweet in context of the inciting event
Analysis
Data traces that document the elements of the original event that the user intends to be public
Offline event
Object
Data

Figure 4: Results of STL decomposition of the time series documenting the diversity of hashtags in use for the period during February 2012 (time scale starts at February 1)

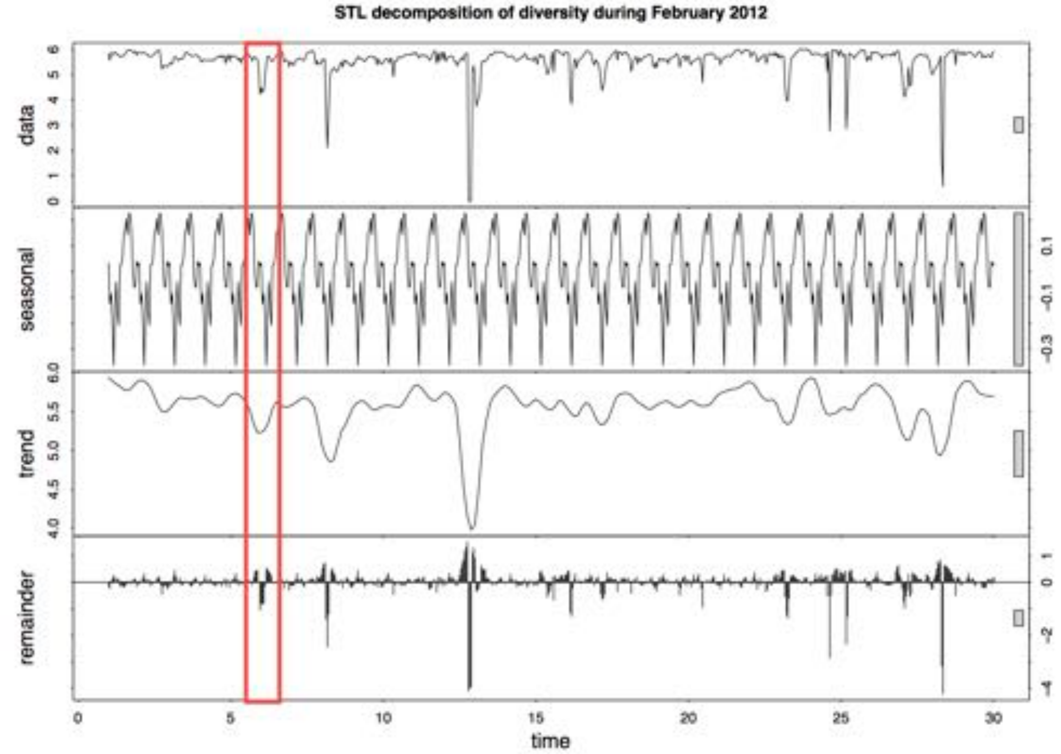


Figure 5: Time series documenting the daily volume of the 1000 top hashtags from February 5th through February 6th. The five hashtags with the highest volume during this time span are identified by colors.

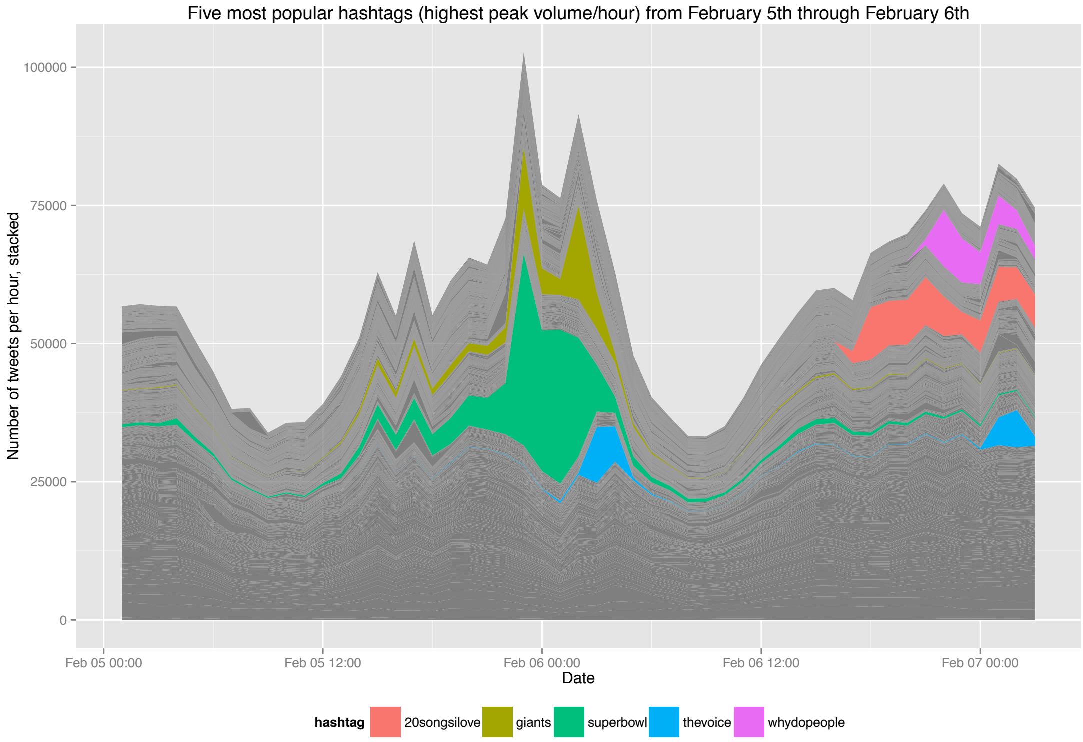


Figure 6: Results of STL decomposition of the time series documenting the diversity of hashtags in use for the period during June 2012 (time scale starts at June 1)

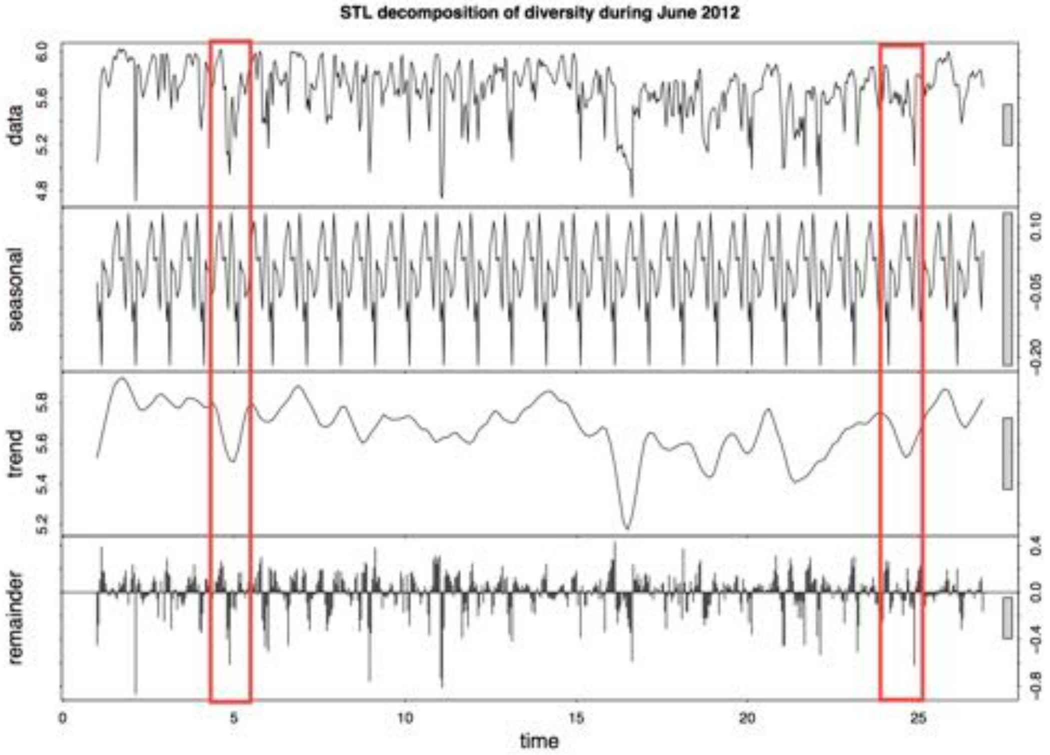
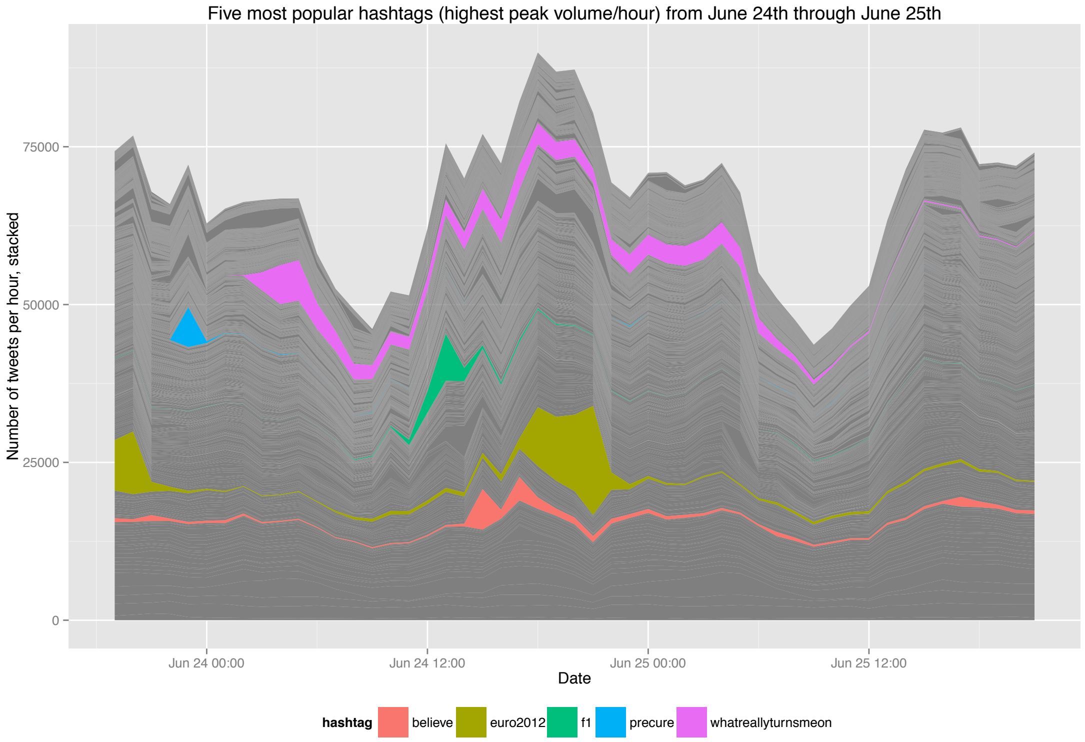


Figure 7: Time series documenting the daily volume of the 1000 top hashtags from June 24th through June 25th. The five hashtags with the highest volume during this time span are identified by colors.


Five most popular hashtags (highest peak volume/hour) from June 24th through June 25th
Number of tweets per hour, stacked
Jun 24 00:00
Jun 24 12:00
Jun 25 00:00
Jun 25 12:00
Date
hashtag
believe
euro2012
f1
precure
whatreallyturnsmeon