Taming Uncertainty in Big Data

Evidence from Social Media in Urban Areas

People's tweeting behavior can be attributed to points of interest in their vicinity. This relationship can be used to identify the veracity of a social media data set. Twitter patterns are recurrent and their stability is an indicator for data certainty. Datasets with high stability estimates can be reliably used in empirical analyses.

DOI 10.1007/s12599-014-0342-4

The Authors

Johannes Bendler, M.Sc. (✉)
Sebastian Wagner, M.Sc.
Dipl.-Vw. Tobias Brandt
Prof. Dr. Dirk Neumann
Information Systems Research
Albert-Ludwigs-Universität Freiburg
Platz der Alten Synagoge
79098 Freiburg i.B.
Germany
johannes.bendler@is.uni-freiburg.de
sebastian.wagner@is.uni-freiburg.de
tobias.brandt@is.uni-freiburg.de
dirk.neumann@is.uni-freiburg.de

Received: 2013-11-05
Accepted: 2014-05-20
Accepted after one revision by the editors of the special focus.
Published online: 2014-08-13

This article is also available in German in print and via http://www.wirtschaftsinformatik.de: Bendler J, Wagner S, Brandt T, Neumann D (2014) Informationsunschärfe in Big Data. Erkenntnisse aus sozialen Medien in Stadtgebieten. WIRTSCHAFTSINFORMATIK. doi: 10.1007/s11576-014-0431-5.

1 Introduction

Big Data, an umbrella term that comprises one of the most promising technologies in today's IT and business world, has been characterized by the four dimensions *volume, velocity, variety*, and *veracity* (IBM 2013). Since it is estimated that 2.3 trillion gigabytes of data will be created every day, *volume* accounts for the sheer amount of data included in Big Data. The second characteristic refers to the *velocity* at which data needs to be processed these days. Social networks like Facebook with more than one billion active users have to handle massive amounts of requests each second. Hence, computer systems need to process large amounts of data in-real or at least near-time. *Variety* comprises the issue of data heterogeneity. Twitter, Instagram and YouTube are just examples of platforms where users are allowed to tweet, post, and upload different formats of data, such as pictures, videos or plain text, often combined with geographical data.

The classic definition of Big Data only included these three characteristics. However, in light of the tremendous growth in social media, IBM added a fourth characteristic, *veracity*, which accounts for the degree of uncertainty in the content of user generated data. The availability of mobile devices enables people to use social media and report subjective opinions and impressions concerning different places, events, objects, as well as other people. This treasure of user-generated data implicitly contains valuable information. Hence, various companies have been collecting social media data in the attempt to understand the needs and desires of their customers and to identify new market trends which may result in innovative product design. No other data source is better suited to understand customers, as users themselves are reporting on their status, activities, dreams and desires. Interestingly, most companies overlook the uncertainty associated with social media: does user-generated data actually contain the meaning assigned to it by analysts? This uncertainty can essentially be characterized according to the three dimensions time, location, and topic. A message may be posted at a specific place and specific time, but it does not necessarily relate to that very place and time. Similarly, topic and inflection of the message may be hard to identify due to semantic and emotional factors.

In this paper we contribute to research on uncertainty by developing a methodology to identify social media data sets with stable causal relationships to time and location of their posting. This methodology enables a fast processing and filtering of social media data to identify relationships that can be used for various purposes, ranging from targeted advertising to policy evaluation.

As a data source for our analysis, we collected almost half a million twitter status messages from within the area of San Francisco including properties such as GPS-coordinates, timestamps, and the text message itself. Additionally, we obtained data on more than sixty thousand points of interest (POIs) including, for instance, restaurants, bars, banks, or museums. We first construct a model that relates a POI to the tweets in its vicinity. This is followed by an investigation of the causal relationship between the POI and the nearby tweets. We thereby validate the model, implying that the tweets around a POI are at least partially caused by the POI itself. Thus, people may possibly not tweet about a particular restaurant (as we do not consider the content of the message), but they are likely to tweet while being at that restaurant and not just randomly passing by. In the final step we investigate recurring patterns within our model. The motivation is clear – if the number of surrounding tweets is causally related to POIs and follows certain daily or weekly cycles for particular

POI categories, these patterns are particularly suitable to be used in further analysis, since they exhibit stronger veracity in temporal and spatial dimensions. For instance, the effect of a specific policy change on bars in a city can be estimated by the deviation of patterns before and after the policy change. Similarly, if the number of tweets around a particular POI category is exceptionally high in the evening, recommender systems should focus on this category at those times. As part of our analysis we construct a statistical indicator based on the jackknifing method that determines the stability of cycles for a given POI category. This indicator helps to quickly identify POI categories for which tweet counts can be reliably used in empirical analyses.

In a nutshell, the research introduced in this paper addresses the following questions:

(1) How can the relationship between a point of interest, for instance a bar or restaurant, and the number of tweets in its vicinity be modeled?
(2) Can a causal link for this relationship be established?
(3) Do some categories show patterns that reliably repeat over time, and can this property be captured in a single indicator?
(4) Which issues can this methodology be applied to in future research?

The remaining paper is structured as follows. The subsequent section presents an overview of research related to our work. In Sect. 3 we model the relationship between Twitter and POI data, thereby addressing the first and second research question. The subsequent analysis of the time series and derivation of the stability indicator as mentioned in the third research question is the topic of Sect. 4. Section 5 concludes with a summary and an outlook on future research.

2 Related Work

In recent years, the amount of data generated every day has reached a level common hardware/software solutions are unable to process within adequate time. Hilbert and López (2011) show – as predicted by Moore's Law – that worldwide technological information processing capacities are exponentially growing. Massive data sets, as created by sensors, cameras, microphones, or radio-frequency identification readers, are very cumbersome to work with, even when using state of the art database management systems and analytic tools. Instead, their evaluation requires software to run parallel on thousands of servers.

This section is divided into two areas of research, which frame our research setting. The first area, dubbed *data uncertainty*, refers to the ambiguities that rest within the data. This part of related work highlights the methodological preliminaries that distinguish our approach from others. The second area, *Big Data in social media*, covers the research stream our approach is applied to. We provide an overview of research on uncertainty in an urban social context and illuminate the main challenges addressed by this work.

2.1 Data Uncertainty

Merging different data streams often reveals high potential in information gain, but at the same time may result in increased uncertainty. Information systems are increasingly required to not only process deterministic events, but also to cope with probabilistic inferences. In this context, Wasserkrug et al. (2005) introduce a framework based on probability theory to represent intertwined events like IT applications and their associated uncertainty. Their seminal work provides a formal treatment to handle event dependent uncertainty in active systems, using dynamically constructed Bayesian networks. In an extension of their research, Wasserkrug et al. (2008) enhance this approach to a mechanism for event materialization under uncertainty. The extended model is again coupled with a Bayesian network and additionally with a Monte Carlo sampling algorithm to account for an accurate probability space of historic events. Their work provides a generic method to handle data under uncertain conditions, but is limited to event-driven architectures.

However, more broadly defined, Big Data Analytics is the process of examining large amounts of data to uncover hidden patterns to improve companies' decision making. Thus, such a restriction to events in the uncertainty modeling is severely limiting. In this respect, Yager (2004) formulates the problem of decision making under uncertainty for decision support. Key aspect of this work is the representation of uncertainty for different situations (in fuzzy sets). Yager emphasizes that decision making requires knowledge of the underlying uncertainty. Inspired by this, our approach provides an indicator that measures uncertainty in social data sets.

Since our research focuses on estimating data veracity in chosen dimensions, which ultimately aims at increasing certainty of inferences that are based on this data, it is useful to differentiate this approach from the traditional IS research stream that attempts to estimate data quality (Heinrich et al. 2007, Boris Otto et al. 2007). While the latter research stream analyzes how data quality can be quantified with respect to properties like completeness, correctness, and timeliness, the veracity approach we follow in this paper aims at measuring uncertainty in the data in order to increase the value of this data. In summary we can state that our approach is novel, as we tackle aspects of veracity in our data, but extend its application beyond systems of events.

2.2 Big Data in Social Media

In the context of social networks, Twitter or Facebook are one major application area Big Data is related to. Ferrari et al. (2011) address the problem of extracting diffuse urban mobility patterns from people's digital life. Accordingly, the authors analyze 13 million tweets in the city of New York and found that the extracted information – albeit noncommittal – can be exploited to identify attractive social hotspots. Furthermore, Wakamiya et al. (2011) attempt to analyze behavioral patterns of crowds in urban areas by using geo-tagged Twitter data in Japan. An experiment shows the ability of social (Big) data to classify areas by common behavioral patterns and categorized them into different types like *bedroom*, *office*, *nightlife*, and *multifunctional* towns.

Directly related to our veracity method using social data patterns Liu et al. (2013) propose a probabilistic factor based recommendation model to determine user preferences for choosing a specific POI. The model takes into consideration geographical influences and the mobility patterns of certain location based social network users. Furthermore, the research of Lee et al. (2011) includes a geo-social event detection method to discover unexpected events by using a large data set of twitter messages. By monitoring specific areas, they show within an experiment that local abnormal crowd activities (Twitter data) can be used to find out expected as well as unexpected events.

Du et al. (2011) have explored the correlation of community and geo information to determine user's mobility behavior in mobile social networks. They compared standard Markov with semi-Markov models and showed by simulation that semi-Markov models are better suited to characterize mobility patterns and more accurate in mobility prediction.

Evidently, existing research on the certainty dimension of Big Data from user-generated sources is still quite limited. This applies even to the context of social media, where uncertainty issues are abundant. Our goal in this paper is not to identify a universal measure for veracity – as the sources of uncertainty vary, veracity can only be attained on a piecemeal basis. Thus, our approach focuses on repetitive occurrence of confirmed causal relationships. In the following Section we will introduce this approach by presenting a model that relates data on Twitter messages to points of interest surrounding the user.

3 Measuring Attractiveness of Points of Interest Using Twitter Messages

One reason for the persisting increase in the daily amount of newly generated data is the pervasive availability of online social services. Facebook, Twitter, Google+, and many others are steadily attracting more users who share messages, images, and other data with the world. These social media services are often used on mobile devices so that their usage is no longer restricted by the confinement of the user's apartment. Instead, users can spontaneously post messages at any location they traverse. A large part of this data is publicly available and may reveal valuable information about user behavior, common interests, or general topics once it has been analyzed. For example, the density of public messages in online social networks along roads or in public transport can help city planners to identify highly frequented routes. The question remains if this data is usable to make decisions on a more finely-grained level. For instance, is it possible to associate the Twitter activity surrounding specific categories of points of interest, such as bars, museums, or restaurants, with those categories?

To clarify this very question is one objective of this work. As mentioned in the Introduction, we develop a method that classifies POI categories according to stable social media usage patterns in their vicinity. However, as a preliminary step, we need to measure the activity of Twitter users in the close surrounding of urban points of interest. This data is subsequently aggregated into the respective POI categories, and used to analyze the attractiveness of those categories over time.

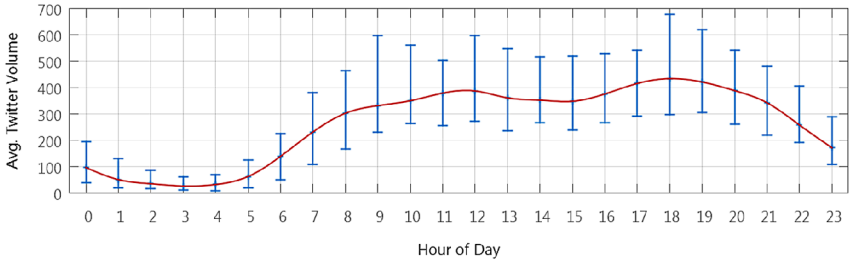

Fig. 1 Observed daily twitter pattern

In the course of this section, our methodology will be introduced as follows. First, we describe our data set for Twitter usage in San Francisco and outline general issues faced when working with data from social media. Second, we present our rank model, which relates individual POIs to the tweets in their vicinity. Thereby, we construct a link between a tweet and the time and location it is sent. The specific determination of ranks and how tweets that are at various distances from a POI are evaluated is explained in the third subsection. As a fourth step, we describe how POI categories are aggregated and ranks vary over time. The final subsection investigates the question of causality and whether changes in tweeting behavior around a POI are causally determined by that specific POI.

3.1 Twitter Usage and Data Characteristics

People follow either their interest or obligation when navigating through a city. During the day, active Twitter users are accustomed to post status messages from time to time in order to inform their family, friends, and followers about their 'everything'. Users are generally tweeting in any mood, but especially when they are really excited or, conversely, in case they are very bored. The latter may astonish at the first sight, but, on the other hand, it is plausible that users kill time by sending status messages that contain rather irrelevant information. Yet, no matter what mood a Twitter user has when posting a message, the geographical location allows to track movement patterns of the user and, more importantly, may implicitly reveal information about the local environment. For instance, we expect locations from which many people are tweeting, to be characterized by some special feature that explains those people's common interest.

For our analysis we have collected all tweets with a geo-location in the municipal area of San Francisco for three months from August 2013 through October 2013. While more than 75 per cent of Twitter users are between 15 and 29 years old (Cheng and Evans 2009), we expect them to be representative with respect to the activity near the majority of POI categories – although some exceptions may apply. Furthermore, more than 35 per cent of all Twitter users post a status message at least once per day.

Our data set reveals an average daily pattern as shown in **Fig. 1**. Starting at a low amount of tweets after midnight, the hourly usage increases in the morning hours and reaches a peak around noon. After a slight drop in the afternoon the tweet count again receives an increase in the evening hours before returning to the night level.

The central reason why we are not able to provide a rapid and especially valid analysis and interpretation of this social data set is uncertainty. Even though many observations over a long period of time may be present, we cannot give precise assertions about data quality, integrity or validity. For social media data in particular, we may encounter uncertainty in any of the three dimensions *time*, *location*, and *topic*.

Concerning *time*, most social networks provide information about the point in time a message was created. However, it is not guaranteed that the message relates to this point in time, as a user can post a message retrospectively. This can also

happen if a user loses his network connection and the mobile device transmits the message later when the connection is up again.

The second dimension of uncertainty is a user's *location* when posting a message. Particular in the age of mobile devices uncertainty with respect to time and to location are strongly related. Some social networks provide a service that either delivers a rough approximation of the urban district, or even transmit the actual geo-coordinates. In most cases, this feature is optional, but often activated by default settings. Again, the location of message transmission does not necessarily tell something about the location the content of the message refers to.

Finally, the factor containing the highest amount of uncertainty in social network messages is their *topic*. In computer science there are various dictionary approaches available that are used to quantify the matter of subjective information, but still semantics are hard to interpret. Other methods use sentiment analyses in order to infer the mood a user had when posting the message. The high amount of abbreviations, creativity in generating emoticons, and incorrect grammar and spelling in online messages further complicates automated content analysis and increases the uncertainty.

These aspects exemplarily show the high degree of uncertainty data gathered from social media suffers from. Depending on data source and application, causes of uncertainty will differ, such that there is no silver bullet to combat uncertainty. In the following approach we, therefore, focus on establishing veracity for the relationship between time and location of tweets and points of interests at those locations.

3.2 Points of Interest Rank Model

The Twitter activity of users in certain districts of a city strongly relates to the density and type of points of interest in the area. Due to the nature of social data, the time of day has a significant impact on user activity around different categories of POIs. For example, we expect 'bars' to indicate a higher activity during night hours than places of to the category 'store'. In order to be able to represent these activity shifts between categories, we introduce an approach that aggregates Twitter activity by POI categories in their close surrounding. Illustrating the terminology, **Fig. 2** shows a randomized example of a map detail from San Francisco. The red dots represent positions of *POIs*, while the blue dots indicate *Twitter messages*. A point of interest can be almost anything that serves as a (publicly accessible) destination for people visiting the district. This includes 92 different categories, such as stores, parks, public buildings, restaurants, or schools. For each point of interest, we define a *neighborhood*, indicated by the black circles. A Twitter message is assigned to a POI whenever its geo-tag is situated within the respective neighborhood.

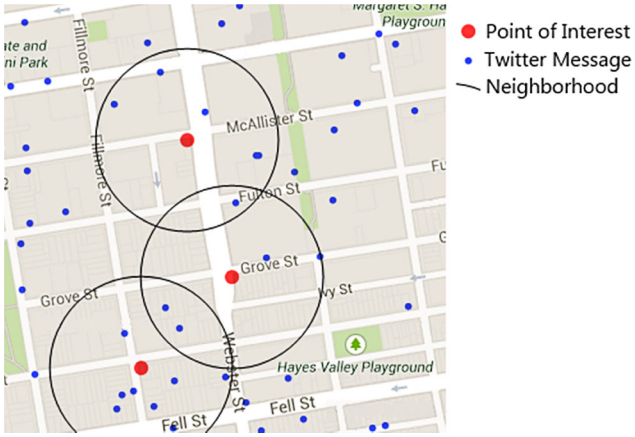

Fig. 2 Randomized example of points of interest

For our analyses and calculations, we use more than 63.000 points of interest as delivered by Google Maps for the city area of San Francisco. The Twitter data at hand has been collected for more than 10 weeks and only consists of geo-tagged tweets, which sum up to almost half a million unique messages.

In context of this research, we define all available points of interest as the set P, where each single element $p \in P$ represents a 3-tuple, defined by Eqs. (1a) and (1b). For each POI, c_p is a set containing the names of all categories the POI is member of, while ϕ_p and λ_p define the GPS-latitude and GPS-longitude values, respectively. The set of all existing categories is defined as C, shown in Eq. (1c) below.

$$P = \{p_1, p_2, \ldots, p_{\|P\|}\} \tag{1a}$$

$$p \in P \mapsto (\phi_p, \lambda_p, c_p) \tag{1b}$$

$$\text{s.t.} \quad c_p \subseteq C$$

$$C = \{c_1, c_2, \ldots, c_{\|C\|}\} \tag{1c}$$

Furthermore, all Twitter messages are comprised within the set T, which is defined in Eq. (2a). Each tweet is represented by a 3-tuple that describes the corresponding details as shown in Eq. (2b). Again, ϕ_t and λ_t refer to the GPS-latitude and longitude values. The attribute τ_t represents the time stamp of the respective tweet, which is the specific point in time where the message was actually published.

$$T = \{t_1, t_2, \ldots, t_{\|T\|}\} \tag{2a}$$

$$t \in T \mapsto (\phi_t, \lambda_t, \tau_t) \tag{2b}$$

3.2.1 Point of Interest Rank Determination

The rank of a single point of interest is a value that reflects the Twitter activity in its neighborhood. In order to calculate an individual rank for a single POI, two equations need to be introduced beforehand. First, we define a function $\sigma(p, d)$ that delivers all tweets from within the corresponding neighborhood.

$$\sigma(p, d) \mapsto N_p \subseteq T, \quad \forall t \in N_p : \text{haversine}(p, t) \leq d \tag{3}$$

Given a reference point of interest p and the neighborhood distance d it yields all tweets N_p with a haversine-based distance smaller or equal to d. The haversine function is commonly used to calculate the spherical distance between two points on the Earth's surface in meters, given their latitude and longitude values. As illustrated by the example in **Fig. 3**, the set N_p consists of all tweets within the black circle around the reference POI (red dot), subject to a radius of $d = 2$.

Furthermore, the relation between Twitter messages and the rank of the respective POI needs to be defined. For this purpose, we define a range-impact function $\gamma(p, t)$ as given in Eq. (4). It delivers a positive real for a reference POI p and a certain tweet t. The resulting value

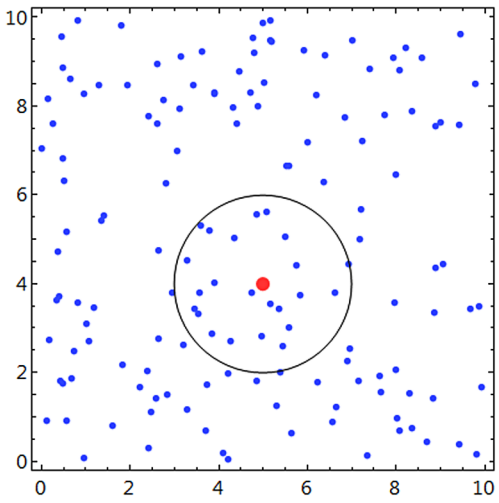

Fig. 3 Point of interest and neighborhood

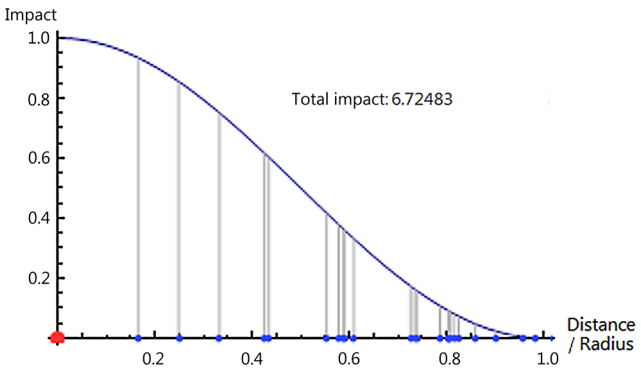

Fig. 4 Range-impact calculation

describes the intensity of the tweet's impact on the point's rank depending on the actual distance between the two.

$$\gamma(p,t) = \frac{1}{2} + \frac{1}{2} \cdot \cos\left(\pi \cdot \frac{\text{haversine}(p,t)}{d}\right) \quad (4)$$
$$\gamma(p,t) \mapsto \mathbb{R}_0^+$$

Referring to the example in **Fig. 3**, **Fig. 4** outlines the calculation of the range-impacts for all tweets from within the neighborhood.

The range-impact function $\gamma(p,t)$ is modeled as a shifted and scaled cosine to fit the range of $[0,1]$ on both x and y-axes. Other possible impact functions could be modeled based on inverse exponentials, linear falloff, or maximum norms. We have decided to choose the cosine shape, due to several reasons. Points of interest are mapped to geo-spherical coordinates by a single point instead of an arbitrary shape, i.e. each data point is described by a pair of latitude and longitude values. On the one hand, referring to the left side of the range-impact plot (cf. **Fig. 4**), tweets that are quite close to the geographical origin of a certain POI can still be seen as closely related to the locations and, thus, should be penalized on a negligible base. On the other hand, tweets that are far from the origin can be ranked with a substantially smaller value but may still be related to the location itself. This means essentially, that the geographical expanse as well as the influence range for potential visitors strongly differs depending on the specific destination. As an example, a football stadium attracts more people from a larger area than an ATM, due to the fact that the number of ATMs within a city is much higher. However, a football stadium spans a substantially larger geographical area than an ATM does. The almost linear fall-off between 'over-penalization' (far distance) and 'under-penalization' (short distance) represents the fuzzy bound of the locations' area. This tradeoff between a higher weight at short distances and a lower weight at far distances complies with the context stated by Tobler (1970) in his first law of geography: 'everything is related to everything else, but near things are more related than distant things'.

The rank r of a single POI is the sum of all impact values of the Twitter messages in the corresponding neighborhood, defined by Eq. (5).

$$r(p) = \sum_{t_k \in \sigma(p,d)} \gamma(p,t_k) \quad (5)$$
$$r(p) \mapsto \mathbb{R}_0^+$$

3.2.2 Time Slicing and POI Category Aggregation

Each point of interest can now be assigned its own individual rank. Towards a complete category rank model, capable of activity analysis by hour of day, two additional changes have to be performed. The original data needs to be split into time slices in order to generate time series and to interpret activity data on an hourly base. Furthermore, the individual ranks of points of interest have to be aggregated by their common categories to allow more general analyses. Therefore, the complete time span of observations is sliced into small time frames τ of one hour length each, such that $\tau_0 \prec \tau_1 \prec \cdots \prec \tau_M$ with $\tau_{i+1} - \tau_i =$ 1h. The time stamps τ_{ti} of tweets have much higher resolution and, thus, can be directly mapped to the time frames τ. Equation (6a) describes the hourly calculation of ranks depending on the time index i. According to this, each POI is assigned an individual rank for each hour from the entire observation time span. Hence, the sum that aggregates the tweets' impacts only includes tweets from within the neighborhood area that have been published in the hour τ_i.

$$r_i(p) = \sum_{t_k \in N_p \mid \tau_i \le \tau_{t_k} < \tau_{i+1}} \left(\gamma(p,t_k)\right) \quad (6a)$$

In a next step, the aggregation by categories is achieved by calculating the average rank over all POIs that belong to the respective category, as outlined in Eq. (6b). The individual ranks within a category are summed up and then divided by the amount of members. Finally, the result is the average rank over all POIs $p_n \in P$ for each category c.

$$r(c \in C) = \frac{\sum_{p_n \in P \mid c_{p_n} = c} (r(p))}{\| p_n \in P \mid c_{p_n} = c \|} \quad (6b)$$

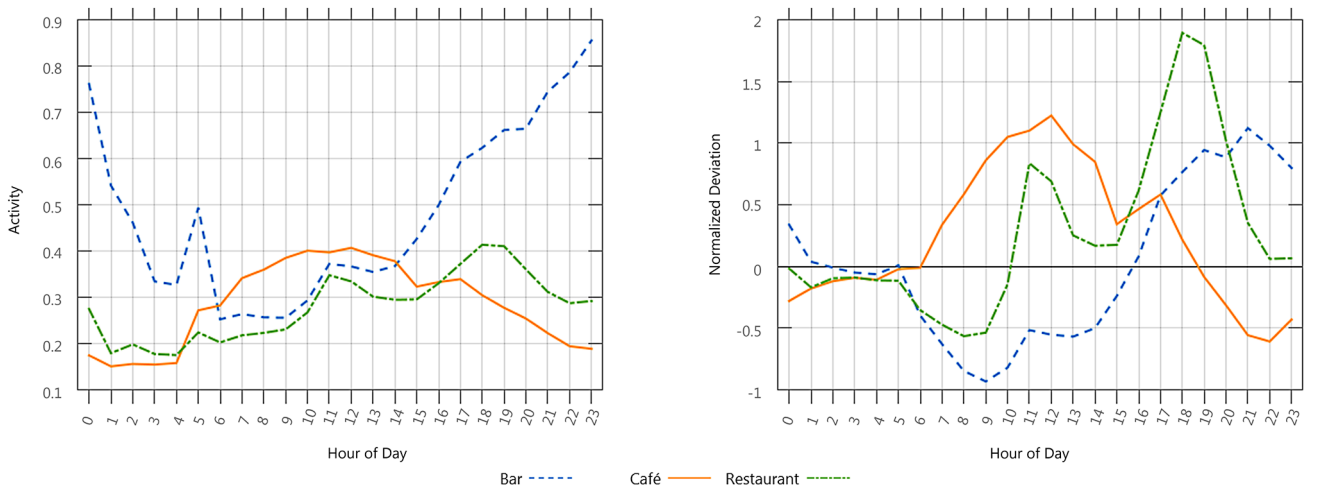

Fig. 5 Pattern comparison for categories 'bar', 'café', 'restaurant'

Table 1 Results of regressing 'bar' rank on total number of messages ('tweets') and bar closing hours ('closed')

Dependent variable	Intercept	Tweets	Closed
bar	1.598 (6.118)	−0.053 (−0.145)	−2.027 (−3.576)

t-Values in parentheses, adjusted R^2: 0.8244

As a last step, Eqs. (6a) and (6b) are applied simultaneously. The rank matrix R is constructed containing a full combination of time frames τ and categories c, as shown in Eq. (7a). This matrix serves as the basis for time series generation and category decomposition in later analyses.

$$R = \begin{matrix} \tau_0 \\ \tau_1 \\ \vdots \\ \tau_M \end{matrix} \begin{bmatrix} r_0(c_1) & r_0(c_2) & \cdots & r_0(c_{\|C\|}) \\ r_1(c_1) & r_1(c_2) & \cdots & r_1(c_{\|C\|}) \\ \vdots & \vdots & \ddots & \vdots \\ r_{\|R\|}(c_1) & r_{\|R\|}(c_2) & \cdots & r_{\|R\|}(c_{\|C\|}) \end{bmatrix} \tag{7a}$$

3.3 Evidence of Causality

The question of causality is central to statistical analysis in general, but even more so to Big Data Analytics. As we often link different variables without a comprehensive theoretical model – in our case the tweets occurring around a POI and the POI itself – , extra measures are required to investigate the causal relationship. After all, one could argue that the rank of a category just depends on the overall number of tweets and not the nature of the category. Hence, the rank of a restaurant for a specific hour might increase simply because more people are tweeting and not because more people are going to and tweeting about that restaurant. We investigate causality in our model on two fronts – visually and analytically.

The visual argument is illustrated in the left panel of **Fig. 5**, where the average ranks for the categories 'bar', 'café', and 'restaurant' are plotted over 24 hours. While 'café' and 'restaurant' apparently loosely follow the average daily tweet pattern given in **Fig. 1**, 'bar' shows the inverse with substantially raised activity during the night hours. Particular to each category, distinct differences can be observed, for instance, 'café' dominates during the morning hours and 'restaurant' slightly peaks in the evening. Moreover, 'bar' shows a spike at 5 a.m. To further investigate these patterns, we normalized all values by subtracting the mean and dividing by the standard deviation, followed by the calculation of the deviations from the normalized total number of tweets. The results are illustrated in the right panel of **Fig. 5** and provide further evidence of category-specific effects. The 'café' rank is exceptionally high – more than would be expected, due to the total number of tweets – in the morning hours and at noon and decreases during the afternoon, albeit with a spike between 4 and 5 p.m. This reflects the typical hours when people visit cafés. The same holds for 'restaurant', which shows positive deviations at noon as well as a very high peak in the early evening hours, and 'bar', which increases during the evening and sustains positive deviations until the early morning hours. The right panel of **Fig. 5** also reveals that the early-morning peak of 'bars' is flattened by using the normalized deviation and thus most likely is subject to the low amount of tweets published during night time.

This visual analysis already provides substantial evidence to suggest category-specific effects, implying that the POI category influences the rank of that point of interest. The question remains if this effect is large enough to be statistically significant, leading to our analytical argument. A common method to investigate causality is through the exogenous variation of a certain covariate as occurring during natural experiments. These have been employed in essentially all evidence-based sciences, including medicine (e.g. Sargent et al. 2004), economics (Zhang and Zhu 2011), and information systems (Kraut et al. 1998). Hence, to estimate the influence of a category on its rank, we need an exogenous variation in the availability of that category. Such a variation is provided for the category 'bar', as it is prohibited to sell alcohol in San Francisco between 2 a.m. and 6 a.m. Therefore, most bars close between 1 a.m. and 2 a.m., some even earlier. We

analyze the effect of these closing hours through a linear regression model, the result of which is summarized in **Table 1**. The normalized 'bar' rank depends on the normalized total number of Twitter messages ('tweets') and a dummy variable representing closing hours ('closed') for the time interval between 9 p.m. and 5 a.m. the following morning. The dummy variable is assigned the value 0 from 9 p.m. to 1 a.m. and the value 1 from 1 a.m. to 5 a.m., as during this time an increasing number of bars are closed.

The regression model works exceptionally well to explain 'bar' ranks, with an adjusted R^2 of 82.44 per cent. For this time interval, the total number of tweets is not significantly correlated with the 'bar' rank. Furthermore, a similar model regressing 'bar' only on 'tweets' provides an adjusted R^2 of just 47.96 per cent. This lack of correlation explains the negative sign of 'tweets' in **Table 1** which, otherwise, would be rather unintuitive. Remarkable on the other hand is the impact of the dummy variable, which essentially drives the explanatory power of the regression model. It expresses the variation in 'bar' ranks that can be attributed to bar closing hours – implying that the presence of this POI category substantially influences its rank.

Both, the visual and the analytical arguments, support the notion that the rank determined for the category 'bar' is not randomly correlated, but instead reliably influenced by people visiting these establishments and that these ranks can be used in future analyses. However, we were lucky to have a good exogenous measure of the availability of bars, a good instrument if you will, to provide analytical evidence of this relationship. To achieve this for all categories seems impossible, since good instruments might not even exist. For instance, 'café' in **Fig. 5** peaks at times when people typically go to cafés, however this is a fuzzy notion and not a good instrument. Nevertheless, we do observe evidence that the ranks of POI categories are likely (partially) driven by causal effects, although these need to be investigated in more depth in future research. However, for the purpose of this research, the results do validate the ranking model in general, such that we can build upon it in the next section.

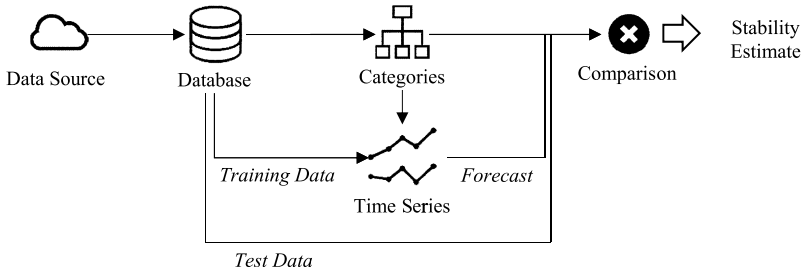

Fig. 6 Pattern stability estimation setup

4 Stability Estimation for Uncertain but Cyclic Data

In this work, we focus on the stability of patterns over time. Without including the content of user-generated messages, we analyze the stability of user behavior in the vicinity of points of interest only based on time and location of tweets. In the previous section we argued that our rank model captures the causal relationship between points of interest and the amount of tweets in its vicinity. This was an important result, as without it the rank model would lack credibility. Our objective is now to derive an indicator that identifies categories of points of interest for which this relationship follows patterns that are stable over time. These categories' patterns obtain credibility, since twitter activity around them is (at least partially) caused by them and follows reliable patterns. Thus, deviations from these patterns are likely to indicate changes in the attractiveness of the underlying POIs.

Consequently, we carry out our calculations for each category of points of interest separately in order to receive an estimated stability value that classifies the steadiness of user behavior for that particular category. The steadier the pattern renders, the less noise is around points of interest of the according category and the more reliable are corresponding user activities. If the measure delivers a low stability value, the automated approach is unable to detect a steady pattern and people are more likely people to just be passing by the point of interest by chance while tweeting.

Figure 6 shows the general setup of our newly proposed methodology. Residing in the Twitter data and POI rank setting, hourly point of interest ranks are collected in a database. For stability of patterns with respect to time we focus on daily and weekly cycles, since our data set does not support seasonal influences. Stability implies in this context that we can compare any two weeks and observe similar patterns. In contrast, instability would result in a time series of category ranks for one week that is completely unrelated to that of another week. To evaluate the stability of any category we designed an approach that is inspired by the statistical method of jackknifing. We use a data sample to train the time series forecast and compare the resulting prediction to a test sample. However, as this would only provide one value that could be prone to outliers and other distortions, we split the training sample into various subsets, which are combined to set up multiple training sets for each category. We use the resulting predictions to measure the consistency in explanatory power for different training sets and, thus, the steadiness and integrity of the categories.

In practice, we worked with ten full weeks w of geo-tagged Twitter status messages from the city area of San Francisco as the data basis for our analyses. The set of all weeks is defined in Eq. (7b). Each week consists of a sub-matrix from the rank matrix R, split among its rows. Since each time slot is of one hour width, every slice results in 168 rows from the original matrix. The set of all weeks is split into a training set A (the first 7 weeks) and a test set B (the final three weeks) according to Eqs. (7d) and (7e).

$$W = \{w_0, w_1, \ldots, w_9\} \tag{7b}$$

$$w_i \in W \mapsto \begin{bmatrix} r_i(c_1) & r_i(c_2) & \cdots & r_i(c_{\|C\|}) \\ r_{i+1}(c_1) & r_{i+1}(c_2) & \cdots & r_{i+1}(c_{\|C\|}) \\ \vdots & \vdots & \ddots & \vdots \\ r_{i+167}(c_1) & r_{i+167}(c_2) & \cdots & r_{i+167}(c_{\|C\|}) \end{bmatrix} \tag{7c}$$

$$A \subset W, \quad A = \{w_0, w_1, \ldots, w_6\} \tag{7d}$$

$$B \subset W, \quad B = \{w_7, w_8, w_9\} \tag{7e}$$

In the next step, we apply partial jackknifing to receive different combinations

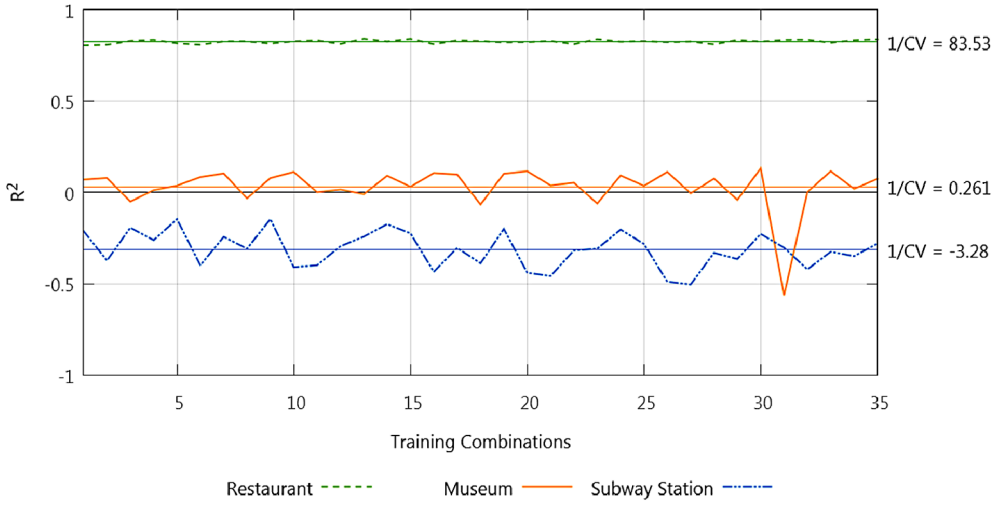

Fig. 7 R^2 distribution characteristics for selected categories

of weeks from the training set as training subsets. This results in 35 combinations with respect to the temporal order: $(1, 2, 3), (1, 2, 4), \ldots, (5, 6, 7)$. For each of these sequences denoted A_i we calculate a three-week-forecast A'_i and compare it to the test set B as defined in Eq. (8).

$$R_i^2 = 1 - \frac{\sum_{k \in A'_i}(B_k - A'_{ik})^2}{\sum_{k \in B}(B_k - \overline{B})^2} \qquad (8)$$

Thus, we generate a set of $35 R^2$-values – one for each training combination – that describe the disparity to the chosen test data category-wise. The R^2 indicator is well established in empirical research to explain the fit of a model by comparing the explained variance to the total variance. While it generally ranges from 0 to 1, negative values are possible when used in our context. As the prediction is based on the training sample and compared to the test sample, the R^2 becomes negative if training and test sample are too dissimilar. However, the implication remains consistent, as a low positive R^2 indicates a bad fit and a negative R^2 an even worse one. We receive a collection of $35 R^2$ values for each category. By analyzing statistical characteristics of each of these collections, we obtain a measure that indicates pattern steadiness. The average R^2 can be high or low, indicated by the mean, and may or may not vary a lot among observations, indicated by the standard deviation. In statistics, a common measure to express the dispersion of a distribution is the coefficient of variation (CV). Applied to the R^2-values, it delivers an estimate of the stability in observed user activity patterns and thus is exceptionally well suited for our demands. Since the mean R^2 may become negative, we calculate the inverse coefficient of variation (ICV) for each category in order to enforce monotonic behavior. Thus, large positive values indicate stable distributions, while negative and small positive values represent the opposite. Since a CV below 5 per cent is generally used to classify stability, we use the inverse value of $\frac{1}{0.05} = 20$ as a threshold representing stable user activity patterns.

Figure 7 exemplary shows the R^2-values for the categories 'restaurant', 'museum', and 'subway station' along with the respective inverse coefficients of variation. The upmost graph represents the category 'restaurant' and visually constitutes a very stable user activity pattern. Its mean is at 0.825, which reveals a strong similarity between the test set and each training combination. On the contrary, the mean of the observations concerning the category 'museum' is at 0.03. Thus, the activity pattern of Twitter users around museums is significantly less stable. Additionally we can identify a high degree of variations, indicating that not all training sets compared equally well to the test data. In even more pronounced contrast, the mean of 'subway station' lies below zero at -0.313. We can infer, that hourly and daily patterns are hardly ever similar for all observations of this category. The visually identified differences are supported by the inverse coefficient of variation values delivered on the figure's right-hand side.

The distribution of all observed categories according to the ICV is shown in **Fig. 8**, where each point refers to a single category. The black line that separates the upper left corner from the rest of the plot represents the threshold where the ICV is equal to 20. Moving to the top and left of it, the ICV increases, while it decreases to the bottom and right. The data points are colored according to their inverse coefficients of variation. Evidently, just about a third of all categories exhibit twitter activity patterns that are consistent over time. Thus, the ranks measured for these categories are reliable and, as we have shown in **Sect. 3.3**, are likely caused by the attractiveness of the respective points of interest. Hence, the stability of these categories' patterns is strong and they can be used for further analyses. In the next section we will discuss what such future research options can be, as well as the limitations to our work.

5 Concluding Remarks

In summary, in this work we developed a methodology to identify online social user interaction in relation to points of interest in an urban setting. Accordingly, we carried out large analyses on nearly half a million Twitter status messages from San Francisco, mapping the hourly activity onto more than sixty thousand points of interest in the vicinity. Based on Tweets published from the immediate neighborhood, we applied a ranking approach to POIs and obtained social activity pattern for each category of POIs. A regression with an instrumental variable approach was conducted that provided evidence for the existence of category-specific effects exemplarily for the category 'bar'. Visual evidence suggested the existence of such effects for other categories, as well. Thus, we can assume and

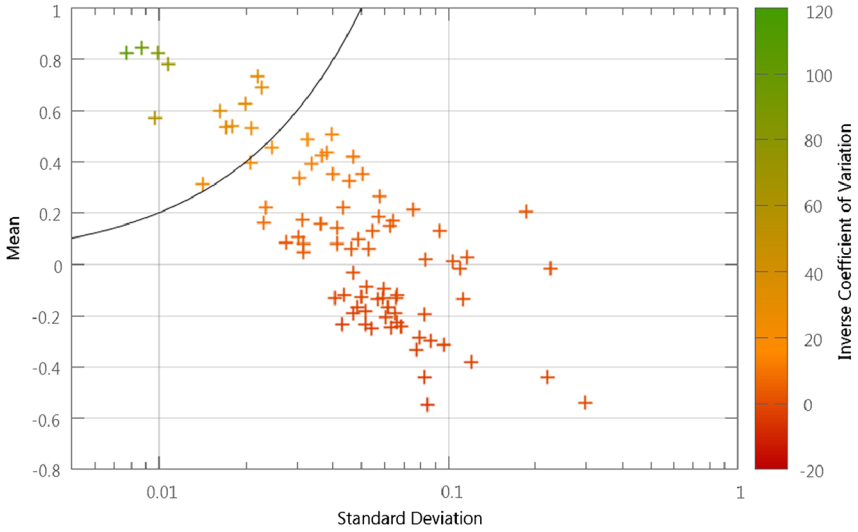

Fig. 8 Distribution of inverse coefficient of variation by category

measure the plausibility of a causal relationship between POI categories and time and location of social activity.

In addition, this work aims at uncertainty reduction when analyzing data that stems from social networks. The results from the previous part show evidence for category-specific patterns concerning user behavior, but still require a measure to indicate reliability. We carried out a statistical analysis inspired by jackknifing and set up time series for the combinatorial space from a 7-week training set for each POI category. The comparison of the time series' forecasts based on the statistical R^2-values was incorporated to develop an indicator describing the stability of user activity patterns. Patterns that recur on a more reliable basis have better prediction accuracy and thus result in a higher stability indication. This pattern stability estimation can be interpreted as a veracity rating according to the temporal and spatial dimensions and serves as a clear-cut metric for certainty even though data may be afflicted with uncertainty and noise. As a consequence of our approach, the uncertain data can be reliably used in empirical analyses.

The two newly developed tools at hand disclose a broad range of application areas and further research. Bolstered by the causal link between POI category and social activity, the category pattern are of special interest and increased value for various purposes besides research, such as recommender systems, targeted advertising, civil protection or city planning. In future research, we plan to use inferred activity by category and hour of day for live-classification of users. Observing Twitter messages from a certain location at a certain point in time, we intend to set up a live recommender system based on the relevant POIs around the location of a user. Additionally, we can extend this research issue by using the estimated stability value from **Sect. 4** for uncertainty modeling. Application of veracity estimation is not restricted to only using the location and timestamp of online social messages. We aim at comprising other metadata, attached web-links, user movement trajectories based on Twitter observations, as well as textual analysis in order to refine our stability estimation. Furthermore, future research can apply the newly introduced methodology no longer only on points of interest but rather extend the calculation to various other reference sets. For example, we aim at performing similar analyses based on local or regional events.

However, the approach suffers from the following shortcomings. The sources of uncertainty are too heterogeneous to capture all aspects of veracity in a single indicator. Although future applications are manifold, the indicator developed in this paper can currently only be applied to social media activity around points of interest – it represents the relationship between a tweet and the dimensions of time and location. Also, the indicator quickly identifies POIs that are likely to have a causal relationship with the surrounding social media activity, which occurs in patterns that are consistent over time. Low values of the indicator, consequently, do not necessarily imply low veracity – it just means that a possible relationship is not consistent over time. One striking example is given by the category 'stadium'. It is very likely that people who tweet from the close surrounding of a stadium are currently attending sports games. Nevertheless, it produced a very low stability estimate. This is plausible, as the game schedule for a single stadium may follow a certain pattern, but the aggregated schedules for all stadiums in San Francisco might not. Thus, future research also needs to investigate further indicators for social media data sets that follow no clear pattern, but are still causally related.

References

Cheng A, Evans M (2009) In-depth look inside the twitter world. http://www.sysomos.com/insidetwitter/

Du Y, Fan J, Chen J (2011) Experimental analysis of user mobility pattern in mobile social networks. In: IEEE Wireless communications and networking conference (WCNC), pp 1086–1090

Ferrari L, Rosi A, Mamei M, Zambonelli F (2011) Extracting urban patterns from location-based social networks. In: Proc of the 3rd ACM SIGSPATIAL international workshop on location-based social networks (LBSN '11). ACM, New York, pp 9–16

Heinrich B, Kaiser M, Klier M (2007) How to measure data quality? A metric-based approach. In: Rivard S, Webster J (eds) Proc of the 28th international conference on information systems (ICIS). Queen's University, Montreal

Hilbert M, López P (2011) The world's technological capacity to store, communicate, and compute information. Science 332(6025):60–65

IBM (2013) The four V's of big data [INFOGRAPHIC]. http://dashburst.com/infographic/big-data-volume-variety-velocity. Accessed 2013-11-05

Kraut RE, Rice RE, Ronald E, Cool C, Fish RS (1998) Varieties of social influence: the role of utility and norms in the sSuccess of a new communication medium. Organization Science 9(4):437–453

Lee R, Wakamiya S, Sumiya K (2011) Discovery of unusual regional social activities using

Abstract

Johannes Bendler, Sebastian Wagner, Tobias Brandt, Dirk Neumann

Taming Uncertainty in Big Data

Evidence from Social Media in Urban Areas

While the classic definition of Big Data included the dimensions volume, velocity, and variety, a fourth dimension, veracity, has recently come to the attention of researchers and practitioners. The increasing amount of user-generated data associated with the rise of social media emphasizes the need for methods to deal with the uncertainty inherent to these data sources. In this paper we address one aspect of uncertainty by developing a new methodology to establish the reliability of user-generated data based upon causal links with recurring patterns. We associate a large data set of geo-tagged Twitter messages in San Francisco with points of interest, such as bars, restaurants, or museums, within the city. This model is validated by causal relationships between a point of interest and the amount of messages in its vicinity. We subsequently analyze the behavior of these messages over time using a jackknifing procedure to identify categories of points of interest that exhibit consistent patterns over time. Ultimately, we condense this analysis into an indicator that gives evidence on the certainty of a data set based on these causal relationships and recurring patterns in temporal and spatial dimensions.

Keywords: Big data, Uncertainty, Social media, Veracity, Spatio-temporal patterns, Points of interest

geo-tagged microblogs. World Wide Web 14(4):321–349

Liu B, Fu Y, Yao Z, Xiong H (2013) Learning geographical preferences for point-of-interest recommendation. In: Proc of the 19th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '13). ACM, pp 1043–1051, New York

Otto B, Wende K, Schmidt A, Osl P (2007) Towards a framework for corporate data quality management. In: ACIS 2007 proc

Sargent RP, Shepard RM, Glantz SA (2004) Reduced incidence of admissions for myocardial infarction associated with public smoking ban: before and after study. British Medical Journal 328:977–980

Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. Economic Geography 46:234–240

Wakamiya S, Lee R, Sumiya K (2011) Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from Twitter. In: Proc of the 3rd ACM SIGSPATIAL international workshop on location-based social networks (LBSN '11). ACM, New York, pp 77–84

Wasserkrug S, Gal A, Etzion O (2005) A model for reasoning with uncertain rules in event composition systems. In: Proc of the 21st conference in uncertainty in artificial intelligence, Edinburgh, Scotland, UAI '05, July 26–29, 2005. AUAI Press, Corvallis, pp 599–608

Wasserkrug S, Gal A, Etzion O, Turchin Y (2008) Complex event processing over uncertain data. In: Proc of the second international conference on distributed event-based systems (DEBS '08). ACM, New York, pp 253–264

Yager RR (2004) Uncertainty modeling and decision support. Reliability Engineering & System Safety 85(1–3):341–354. doi:10.1016/j.ress.2004.03.022

Zhang X, Zhu F (2011) Group size and incentives to contribute: a natural experiment at Chinese wikipedia. The American Economic Review 101(4):1601–1615