

The emergence of complex patterns in online human communication

Joachim Mathiesen*, Pernilly Yde, Mogens H. Jensen

Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, DK-2100 Copenhagen, Denmark

* E-mail: mathies@nbi.dk

Abstract

Social media have become essential conduits in the worldwide exchange of ideas, opinions and consumer marketing. Complex networks are important tools for analyzing the information flow in many aspects of nature and human society. Here, we introduce a method based on networks and social media to gauge how ideas, opinions and new trends impact society. We show that correlations between different international brands, nouns or US major cities follow a universal scale free distribution. The correlations indicate a self-organizing dynamics in large social organizations where the exchange of information between individuals is highly volatile. Our method provides new fundamental insight on the propagation of opinions and the emergence of trends in online communities.

Introduction

Networks are elegant representations of interactions between individuals in large communities and organizations [1–5]. These networks are constantly changing according to demands, fashions and flow of ideas [6, 7]. Recently there has been a growing interest in the dynamics of complex networks with a focus on pairwise interactions [8–10]. However, it has often been assumed that the mere existence of a connection in a network implies that transmission of information is complete. This assumption is associated with the fact that most studies consider static structures derived from all recorded interactions. Often it is less interesting though to know whether two individuals are connected at some point in time and space than it is to quantify the volatility of the link connecting them.

The rapidly growing flux of information through online media permits an unprecedented analysis of human behavior and interactions [11, 12]. These interactions can be monitored in real time with a high level of detail via social media such as Twitter [6, 7]. Twitter is a micro-blogging universe where registered users can submit small pieces of information, named “tweets”, to an online stream. The length of a tweet is limited to 140 characters and the content ranges from personal information to massively distributed advertisements or political messages. Twitter has a potentially huge reach and is used by an increasing number of companies and political organizations to disseminate news.

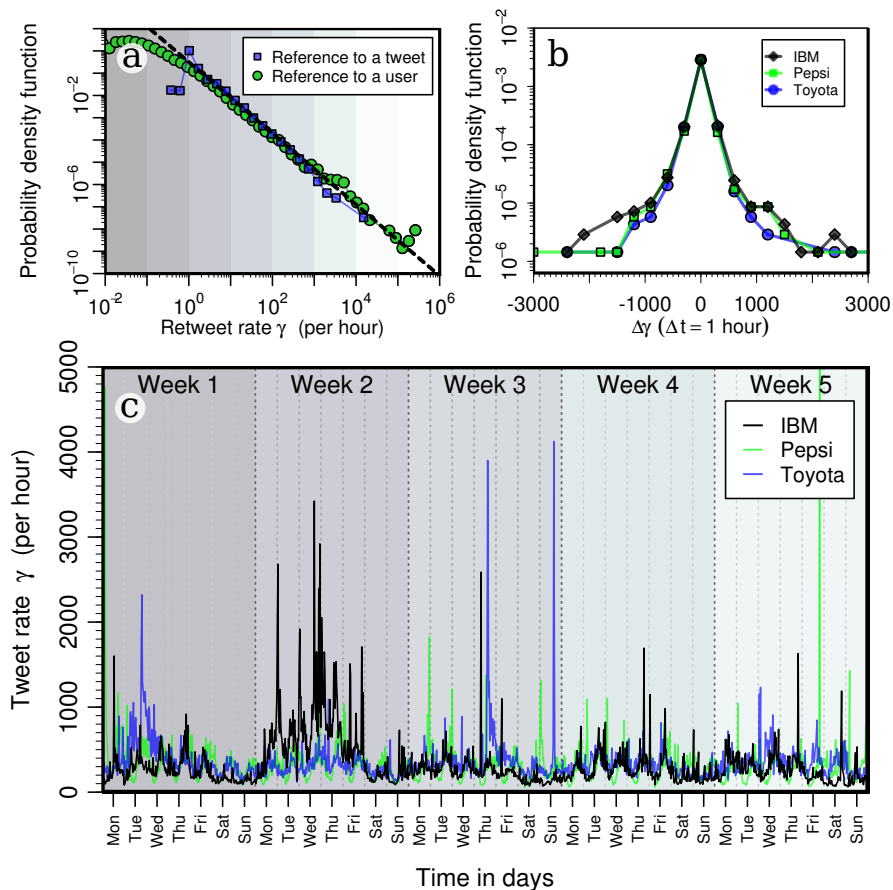


Figure 1. Intermittent dynamics of interacting users of social media. Panel a), probability distribution function of reference rates to individual tweets (blue squares) and users (green circles). The distribution is presented using double logarithmic axes. The dashed line is a best fit to the scale free distribution with an exponent $\alpha = -1.70 \pm .05$ (s.d.). Panel b) and c) show the temporal variation in the tweet rates of three international brands. In the time signal in panel c) there is a clear intermittent fluctuation in the overall signal while at the same time there is an underlying periodic variation over days and weeks. Panel b) shows the corresponding distributions of the tweet rate change $\Delta\gamma_t$ measured in hourly intervals.

Results

Current tweet rates were measured by submitting repeated search queries to Twitter. For each query, up to 1500 of the latest tweets were returned and based on the time interval over which they appeared a rate was calculated. Samples were performed during a 4 month period, November 2010 to February 2011. Users of Twitter have the opportunity

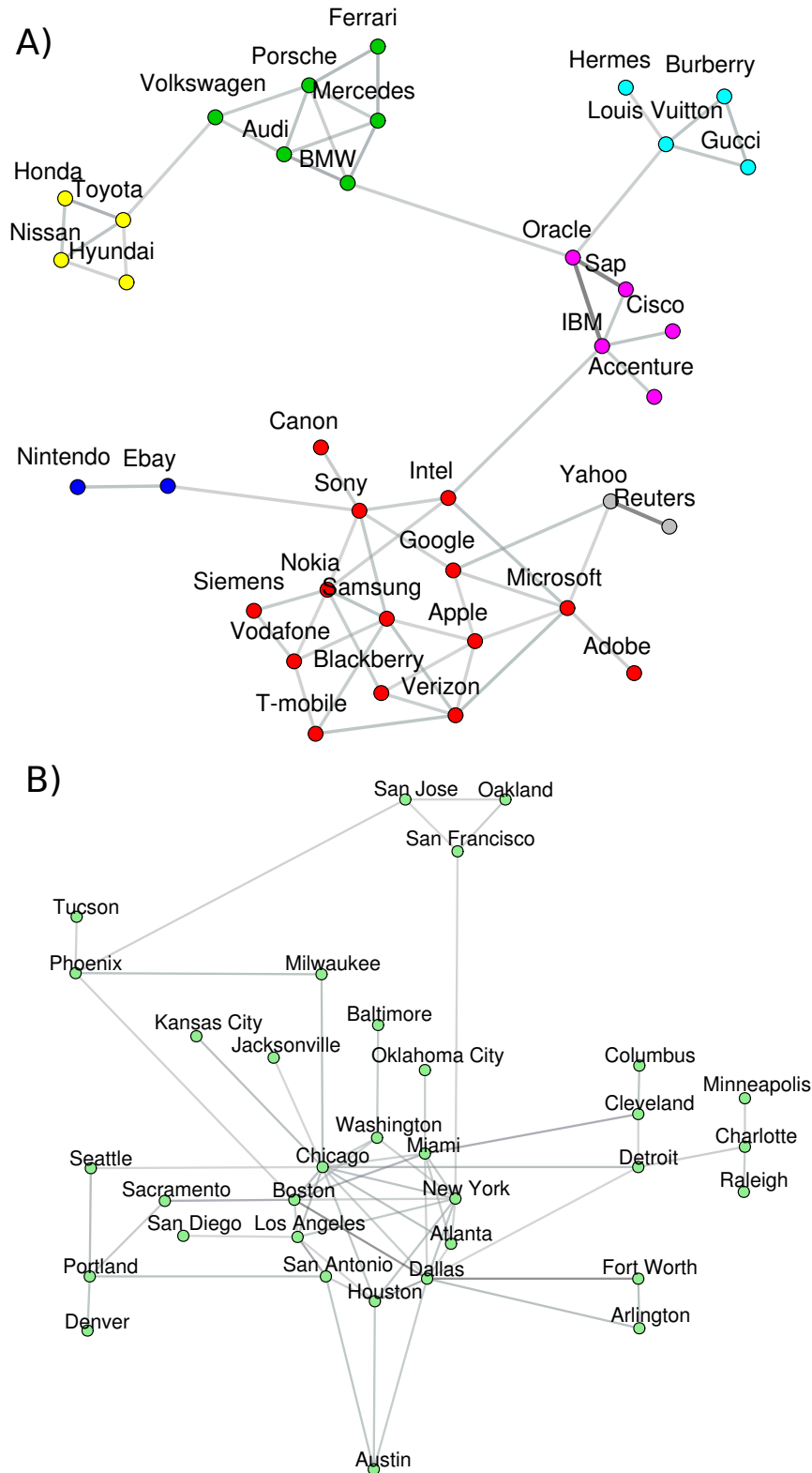


Figure 2. Network of correlations based on tweet-rates on Twitter.

Correlations between international brands are shown by the network in Panel A). A link in the network represents a similarity measure between brands computed using Eq. (1).

Only links with a strength larger than 0.004 are shown. The network is strongly modular and individual modules have been identified using the spin-glass method [19].

In panel B), a similar network over major cities in the USA is shown. Links with a

to actively follow and re-post tweets, so-called retweets, of other users. The rate by which a given tweet or user is retweeted provides an instantaneous measure on the community's interest in a given subject. The distribution $p(\gamma)$ of retweet rates γ of specific tweets and users are shown in Fig. 1. Both distributions have a scale free form on up to five orders of magnitude with an exponent 1.7 ± 0.1 (s.d.), i.e. $p(\gamma) \propto \gamma^{-1.7}$. The scale free distribution strongly indicates that the community's interest in a given user or subject is self-organizing [13–16] and as a consequence leads to avalanches of retweet rates of all sizes. Avalanches are often associated with an intermittent dynamics where large to extreme events are interrupted by longer periods of quiescent behavior. This is confirmed by the distribution of the tweet rate change $\Delta\gamma_t$ measured in hourly intervals. The distribution has a clear stretched exponential tail Fig. 1b. Similar 'heavy-tail' distributions are observed in a number of other systems which show intermittent temporal dynamics. Examples are fully developed turbulence and time series of economical indices [17, 18]. The impact of a current subject is influenced by the collective behavior of individuals in social networks. For instance, the constant stream of information can be used to monitor the real-time popularity of different topics. Moreover, correlations between topics can be measured by analyzing the content of individual tweets, i.e. related topics are likely to appear simultaneously in a tweet as well as in associated retweets. Specifically, we have analyzed correlations within three widely different categories: 1) international brands, 2) nouns and 3) major cities in the USA. We compiled a list of 100 popular international brands representing top companies in different categories and used a list of the 50 largest cities in the USA. Similarly random samples were taken from a list of 2000 common nouns of the English language. The similarity of two entities A and B is defined in terms of the rate γ_{AB} by which tweets contain both A and B . For example, by considering queries to Twitter containing the terms "Google" and "Microsoft", we get $\gamma_{Google} \approx 130000$ per hour and $\gamma_{Microsoft} \approx 17000$ per hour whereas $\gamma_{Google,Microsoft} \approx 700$ per hour (January 2011). A normalized symmetric measure of similarity is naturally defined by

$$\omega_{AB} = \frac{\gamma_{A \cap B}}{\gamma_{A \cup B}} = \frac{\gamma_{AB}}{\gamma_A + \gamma_B - \gamma_{AB}} \quad (1)$$

In Fig. 2 we present networks of international brands and USA major cities created by this measure. The network of brands is strongly modular with groups of brands representing similar products. However, some links reveal non-trivial relations between selected brands. For cities the similarity network provides an alternative map where individual cities only to some extent are grouped according to their geographical location. The network is dominated by a central module consisting of New York, Chicago, Atlanta, Los Angeles and Boston. This is not surprising as these cities are hubs in the American society. In the lower right part, we observe Californian cities in a module that connects naturally to cities like Denver and Seattle. We also detect a module of east-coast to mid-western cities connecting to a module of southern cities.

In social media, international brands and to some extent US cities are of global common interest. Fashions spontaneously emerge over a short time span in terms of a collective

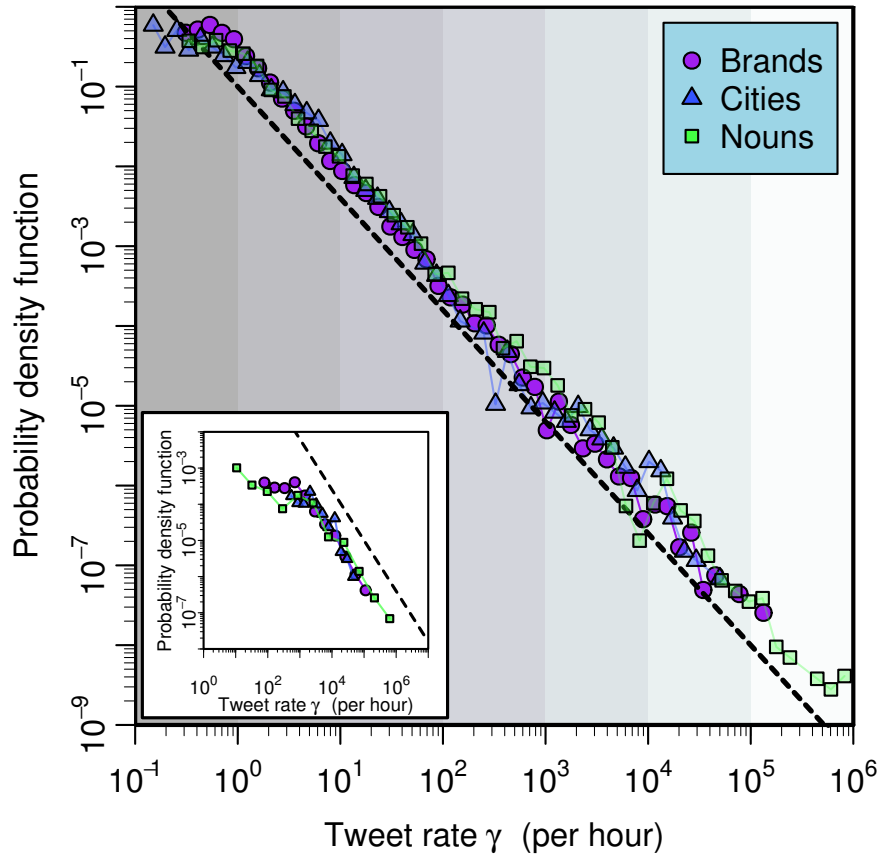


Figure 3. Probability density function of tweet rates of pairs of international brands, major cities in the USA and common English nouns. The distributions include rates of individual search terms. The violet circles correspond to brands, the blue triangles to cities and the green squares to nouns. Note that the rates of the cities have been multiplied by 20 to allow for a direct comparison. The data (measured per hour) were obtained by repeated queries to Twitter of pair-wise terms in the period November 2010 and to February 2011. The distributions of the rates are scale invariant over more than six orders of magnitude and have the same exponent $\alpha = -1.40 \pm .02$ (s.d.). The dashed line is a guide to the eye and corresponds to the computed exponent. The inset shows distributions of tweet rates of single brands (purple circles), major US cities (blue triangles) and English nouns (green squares). Although there is no clear scaling of these distributions we have inserted the same line as in the main panel for comparison.

awareness in the network. We therefore expect that correlations between current fashions reflect how the awareness percolates on the social network. As a main result, we obtain scale free distributions of the pair-wise tweet rates γ_{AB} over six orders of magnitude using brands, nouns as well as major cities, see Fig. 3. Surprisingly, the distributions are all defined by the same scaling exponent 1.40 ± 0.02 (s.d.). The distribution of the tweet rates of individual search terms A , γ_A , does not follow a clear scale invariant distribution (see inset of Fig. 3) and correlations cause the tweet-rate of pairs γ_{AB} not to be proportional to the product $\gamma_A \gamma_B$. In particular we notice that if the distribution of the rates γ_X could be approximated by a scale invariant distribution $p(\gamma_X) \sim \gamma_X^{-\alpha}$, we get that the product $Z = \gamma_A \gamma_B$ would follow a distribution

$$p(Z) \sim Z^{-\alpha} \log(Z^2). \quad (2)$$

The logarithmic correction to scaling does not fit the data of Fig. 3, e.g. a best fit gives an exponent $\alpha \approx -2$ significantly larger than expected from the distributions of the tweet rates γ_X of individual search terms shown in the inset of Fig. 3. We attribute this observation to the interactions between individuals of the social network which generate a common perception of related entities. In other words, there is an amplification of the frequency by which e.g. certain brands appear together in a tweet. Performing a similar analysis using search engines such as Google and Bing, we achieve different distributions (see Fig. 4a). In contradistinction to social media, the search engines integrate over a long time and include results from widely different media. Moreover the search engines include results from web pages which are not restricted in size like the tweets.

Discussion

The results can be put into further perspective by considering correlations of nouns in a different context. In sentences of novels by e.g. Mark Twain (Huckleberry Finn) and Herman Melville (Moby-Dick), we recover scale free distributions with significantly larger exponents (see Fig. 4b). The sentences have a typical length comparable to the 140 characters of a tweet. However, a novel is written by a single author and typically exhibits a more formal structure compared to the text messages created by online interactions between many individuals.

The unique self-organizing behavior of users of social media appears to initiate a cascade dynamics which widens the distributions and lowers the scaling exponent relative to that of novels. A deeper understanding of this effect calls for an uncovering of mechanisms behind human communication on online media. Social media have become vital channels for advertising as well as disseminating news and political opinions, therefore this understanding will have significant potential not only in several branches of science but also for commercial purposes. The fact that the complex communication patterns appear to be universal indicates that the information flow on social media is a reflection of basic human behaviour.

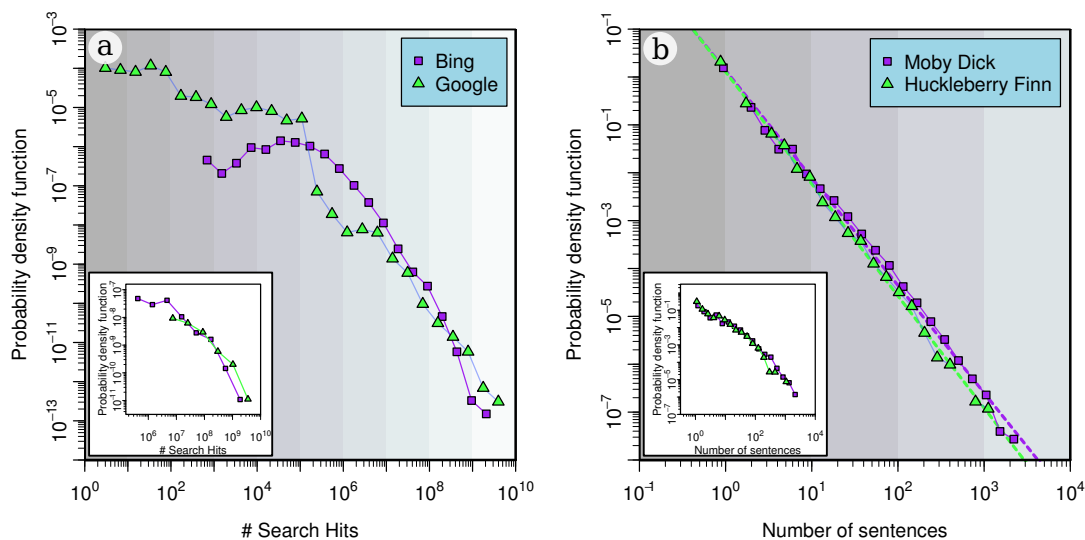


Figure 4. Probability density functions of the number of search hits returned from Bing and Google and for the number of sentences in which two nouns co-appear in novels. In panel a) we performed pair wise queries on international brands to Bing and Google. In contrast to the result obtained from Twitter, we do not observe clear scale-free distributions. Inset: Probability density functions of search hits returned from queries on individual brands alone. Panel b) shows the number of sentences in which two nouns co-appear in the novels Huckleberry Finn (Mark Twain) and Moby-Dick (Herman Melville). The distributions are plotted on double-logarithmic scales and include the distributions of individual nouns. Dashed lines are best fit to a scale-free distribution and have exponents $-2.34 \pm 0.04(\text{s.d.})$ (Huckleberry Finn) and $-2.24 \pm 0.04(\text{s.d.})$ (Moby-Dick). Inset: Probability density function of the frequencies by which individual nouns appear in the same sentences.

Acknowledgments

Suggestions and comments by Kim Sneppen and Namiko Mitarai are gratefully acknowledged. This study was supported by the Danish National Research Foundation through the Center for Models of Life.

References

1. Albert R, Barabasi A.-L. (2002) Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74:, 47-97.
2. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. *Phys. Rep.* 424: 175-308 (2006).
3. Borgatti SP, Mehra A, Brass DJ, Labianca G. (2009) Network analysis in the Social Sciences. *Science* 323: 892-895.
4. Kitsak M *et al.* (2010) Identification of influential spreaders in complex networks. *Nature Physics* 6:, 888-893.
5. Vespignani A (2010) Complex networks: The fragility of interdependency. *Nature* 464: 984-985.
6. Mandavilli A (2011) Peer review: Trial by Twitter. *Nature* 469:, 286-287.
7. Huberman BA, Romero DM, Wu F (2009) Crowdsourcing, attention and productivity. *J. Inform. Sci.* 35: 758-765.
8. Newman, MEJ (2010) *Networks: An Introduction*, Oxford University Press, New York.
9. Nagler J, Levina A, Timme M (2011) Impact of single links in competitive percolation. *Nature Physics* 7:, 265-270.
10. Buldyrev SV, Parshani R, Paul G, Stanley HE, Havlin S (2010) Catastrophic cascade of failures in interdependent networks. *Nature* 464: 1025-1028.
11. King G (2011) Ensuring the data-rich future of the social sciences. *Science* 331:, 719-721.
12. Centola D (2010) The spread of behavior in an online social network experiment. *Science* 329: 1194-1197.
13. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: An explanation of the 1/f noise. *Phys. Rev. Lett* 59: 381-384.

14. Ramos O, Altshuler E, Maaloy KJ (2009) Avalanche prediction in a self-organized pile of beads. *Phys. Rev. Lett* 102, 078701.
15. Jamtveit B, Jettestuen E, Mathiesen J (2009) Scaling properties of European research units. *Proc. Natl. Acad. Sci. USA* 106: 13160-13163.
16. Mathiesen J, Jamtveit B, Sneppen K (2010) Organizational structure and communication networks in a university environment. *Phys. Rev. E* 82: 016104.
17. Bohr T, Jensen MH, Paladin G, Vulpiani A (1998) *Dynamical system approach to turbulence*, Cambridge: Cambridge University press.
18. Mantegna RN, Stanley HE (1995) Scaling behaviour in the dynamics of an economic index. *Nature* 376: 46.
19. Reichardt J, Bornholdt S (2006) Statistical mechanics of community detection. *Phys. Rev. E* 74: 016110.