

Word usage mirrors community structure in the online social network Twitter

John Bryden¹, Sebastian Funk^{2,3*} and Vincent AA Jansen¹

*Correspondence:

sf7@princeton.edu

²Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA

³London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

Full list of author information is available at the end of the article

Abstract

Background: Language has functions that transcend the transmission of information and varies with social context. To find out how language and social network structure interlink, we studied communication on Twitter, a broadly-used online messaging service.

Results: We show that the network emerging from user communication can be structured into a hierarchy of communities, and that the frequencies of words used within those communities closely replicate this pattern. Consequently, communities can be characterised by their most significantly used words. The words used by an individual user, in turn, can be used to predict the community of which that user is a member.

Conclusions: This indicates a relationship between human language and social networks, and suggests that the study of online communication offers vast potential for understanding the fabric of human society. Our approach can be used for enriching community detection with word analysis, which provides the ability to automate the classification of communities in social networks and identify emerging social groups.

Background

The complexity and depth of our language is a unique and defining feature of humans. Language permeates our daily lives as we use it to convey information from simple messages to opinions and complex arguments. In addition, it has a number of functions that transcend the transmission of information, with a range of social implications. Sociolinguistic studies have shown how varieties of a language can be strongly associated with established social or cultural groups [1–5]. In general, these studies have tended to concentrate on small, distinct and relatively stable communities such as gangs [6, 7] or inner-city working communities [8].

In the study of complex networks, the term *communities* is used to denote parts of the network that are more strongly linked within themselves than to the rest of the network, a phenomenon that has been observed in many human social networks [9]. In this sense, communities are an emergent property of network structure. Much work has gone into developing methods to detect such groups from topological analysis [10], and the extent to which this is possible has been termed *modularity* [11]. The communities found in this way are usually associated with groups of friends or acquaintances, or similarity in traits [9, 12, 13]. If these communities overlap with social or cultural groups, the use of language

should vary between different communities in a social network [3]. Taking word usage as a proxy for variation in language [14, 15], we hypothesise that this variation should closely match the community structure of the network.

To test this hypothesis, we studied word usage in a weighted network created from communication between about 250,000 users of the social networking and microblogging site Twitter, and analysed if groups identified within the interaction network indeed had unique language features. Twitter communication is unstructured in the sense that every user can send a message to any other user. In constructing our network, we formed a link only when users had mutually directed messages at each other, analogously to what has been done in the study of mobile phone networks [16]. We used methods from statistical physics and network theory to identify groups in the network structure that emerge from user interaction, and linked this to word frequencies in the messages generated by each user.

Results and discussion

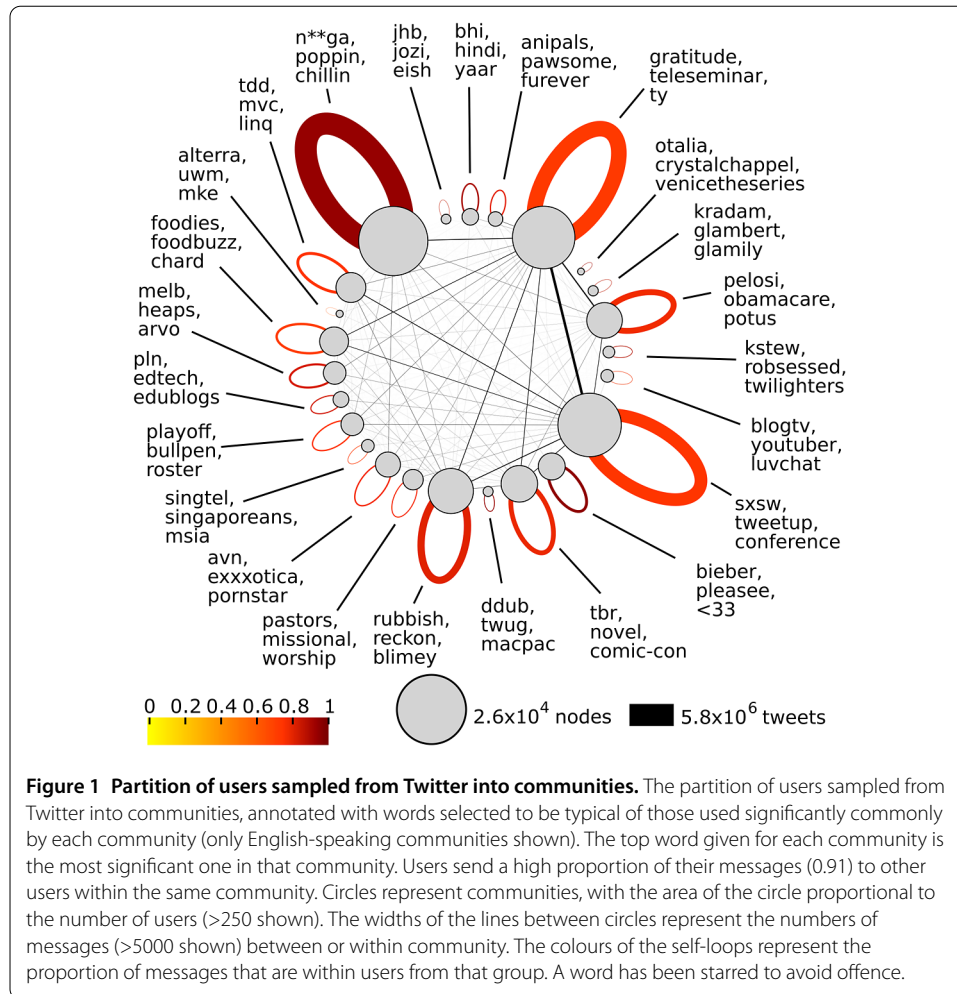
The network analysed had 189,000 nodes (each corresponding to a single user) with 75 million mutual tweets between them (mean degree of 28) and a global clustering coefficient of 0.084.

Characterising communities through word usage

Partitioning the sampled network of twitter users into groups so that the proportion of messages between users of the same group was maximised [17] yielded pronounced community structure (Figure 1). The modularity found with this *maximum modularity* algorithm is $Q = 0.78$, in the order of previously studied social networks [17] and far greater than the expected maximum modularity for a random network of the same degree distribution ($Q_{\text{random}} = 0.15$) [18]. For the larger communities identified, the algorithm we used identified sub-communities forming a hierarchy of communities (modularity at the lowest level was $Q = 0.66$). Testing the network for community structure using a second algorithm which uses the *map equation* to partition users into groups such that flow in the network stays within groups [19–21] yielded modularity $Q = 0.67$. We focus on the high-level partition generated by the modularity maximisation algorithm, but will return to the map equation algorithm to verify results.

We characterised each of the communities according to the words used in messages sent by the users of the community. To do this, we ranked words in each community by the Z-score of their usage to identify the words most representative of that community. Figure 1 gives illustrative examples of words that characterise each English-speaking community of more than 250 users (see Additional file 1 for the lists of top-ranked words). We surveyed the mean global frequencies for the 100 top-ranked words of each community, finding a broad range. Some communities used relatively common words (at 13% of global usage), while others used much rarer words (at 0.04% of global usage).

To determine the significance of word usage differences, we calculated the Euclidean distance of relative word usage frequencies for each pair of English-language communities using a bootstrap. For each such pair of communities, we sampled two new groups (with replacement) from their union until they had the same sizes as the communities being compared. Repeating this procedure 1,000 times for each pair of communities, we found that for 248 of the 253 pairs of communities the distance between the original pair



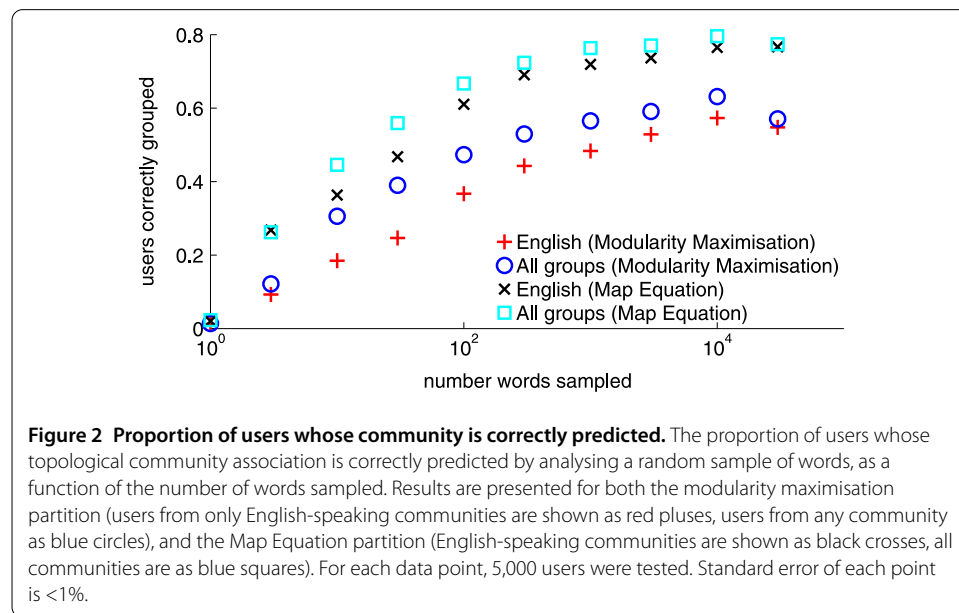
was greater than all of the 1,000 resampled pairs. For the other five pairs of communities, the distance between them was greater than most ($\geq 95\%$) of the resamples. Comparable results were found for the communities generated with the hierarchical map equation algorithm. In other words, the community membership can explain part of the variance of word usage.

One could argue that the within-community similarities between word frequencies arise because users within a group communicate about one or a limited number of subjects. In contrast, there are revealing differences in word usage that go beyond subject area (see Table 1). To quantify differences beyond subjects of conversation, we tested other statistics using the same method we used for testing differences in word usage: frequency of letter usage, letter pair usage, word length and three-letter word ending. For all these statistics, the distances between almost all the community pairs ($>98.8\%$) was greater than the distances between 95% of the 1,000 pairs of groups resampled from the unions of communities. The best-performing statistics were word ending and letter-pair frequency, and the worst was word-length frequency. Only a very small fraction of these matching word features ($<10\%$) came from the same words. These results suggest that the communities used different language patterns, even when considering quite subtle differences that go beyond common subjects of conversation, such as word endings or word lengths.

Table 1 Language patterns of communities

Community	Language feature	Number of occurrences
n**ga, poppin, chillin	shortened endings ('er' → 'a' or 'ing' → 'in')	50
pln, edtech, edublogs	amalgamated words	31
anipals, pawsome, furever	animal based puns	31
bieber, pleasee, <33	lengthened endings (repeated last letter)	28
kstew, robsessed, twilighters	amalgamations/puns around Twilight movie genre	28
tdd, mvc, linq	acronyms	25
kradam, glambert, glamily	puns around pop star Adam Lambert	15

Language patterns found in communities from Figure 1. We looked at the top 100 words used significantly more than the rest of the population for each community, identified language patterns, and counted the number of occurrences. A word has been starred to avoid offence.



Predicting community membership from word usage

We also tested whether individual word usage can predict the community membership of users by comparing their own word frequencies with community word usage (see Materials and Methods). The more words we sampled from a test user, the more likely we were to associate the same community with the user as that which was found from topological community detection (Figure 2).

Comparing different partitions, the hierarchical map equation predicted the community of more users correctly than the high-level modularity maximisation partition. Analysing a random sample of 1,000 words of each user, we were able to predict the correct map equation community for approximately 72% of English-speaking users, compared to 48% with the modularity maximisation partition. Using the numbers of communities generated by each algorithm ($N = 322$ for the map equation versus $N = 413$ for modularity maximisation), we calculated the Z -score for these prediction scores. The Z -score for the map equation ($Z = 911$) was greater than that for the modularity maximisation ($Z = 687$). When, on the other hand, considering the lowest-level partition produced by the modularity maximisation algorithm, the fraction of users predicted correctly drops to 38%. When taking into account the number of communities ($N = 8,738$), though, the Z -score is greater than for both of the other partitions ($Z = 2,511$).

Given the community structure of the network, around half of messages will be directed to users in the high-level communities we predict. This means that, once a network is analysed, it is possible to assign the most likely community or communities for any user that was not part of the community detection. This can be done solely on the basis of the word frequencies in a relatively small sample of text written by that user. The proportion of topological groups predicted correctly from analysis of word usage increases roughly exponentially with the number of words sampled from each user (Figure 2).

Conclusions

We studied the relation between community structure in an online social network and language use in messages within that network, and found a striking overlap, whether we considered words, word fragments or word lengths. Moreover, we were able to predict the network community of a user, a purely structural feature, by studying his or her word usage, and we found that this was possible with rapidly growing accuracy for relatively few words sampled. This indicates how the language we use bears the signature of societal structure, and is suggestive of the enormous potential in using topological analysis to identify cultural groups.

A pair of users that engage in an online conversation would be expected to have some language in common. When groups of individuals share language, and also converse with each other, then it is possible to use our method to identify these groups and enrich them with the language they are using. A wide range of alternative algorithms may also be used [22]. A full exploration of these is beyond the scope of this paper, but may show improvements in identifying communities with more unique language patterns. Further improvements might be made by replacing Z -score metric we have used to identify words that stand out with a term frequency-inverse document frequency metric [23].

Our sample is only a small proportion of the much larger twitter network and one could ask whether the sampling process introduces a bias in the community structure we detect. Our sample network has small-world properties (average shortest path length $L = 4.4$), indicating that the sampling process should very quickly reach every community in the network. Resampling the network confirms this intuition. For very small resamples (<1%), the shortest path length L_r is greater than L , but on further sampling it converges towards L . Similarly, modularity decreases initially with the size of the resample before it converges, indicating that after enough sampling the process is no longer biased toward any particular community. This is consistent with previous analysis of this type of sampling process which showed that (given certain assumptions) it is a regular Markov process [24], and thus that the community being sampled is independent of the community at the origin of the sample [25]. Overall, this resampling analysis demonstrates that our sampling procedure quickly discovers the larger communities in the network if they are not completely isolated. With more sampling, smaller communities and sub-communities can also be discovered.

The finding that people can be placed in a community by analysing their language usage is consistent with evidence that humans make long-term decisions about relationships very quickly [26]. Our results give an indication that words could be markers of desirable underlying traits or social norms [27], allowing people to make quick decisions about the type of relationship they want from a new acquaintance. The community structure we observe in the network could be explained through homophily [13, 28], that is, through

people biasing their interactions to others that are similar in some way, or through dyadic interactions [16]. More generally, any process that structures people into groups could play a strong role in cultural evolution [29–32], as well as in the spread of information or pathogens [33, 34]. If people with a negative attitude towards vaccination are preferentially in contact with those of the same opinion, this could lead to clusters of susceptibles and increased risk of outbreaks [35]. There is clearly scope for further study of the role such structuring plays in the evolution of cooperation in humans [36].

Online social networks offer us an unprecedented opportunity to systematically study the large-scale structure of human interactions [37]. Our approach suggests that groups with distinctive cultural characteristics or common interests can be discovered by identifying communities in interaction networks purely on the basis of topological structure. This approach has several benefits when compared to surveying groups identified on a smaller scale: it is systematic, and groups are identified and classified in an unbiased way; when applied to online social networks it is non-intrusive; and it easily makes use a large volume of rich data. In this study we characterise groups by their word frequencies, but this could be extended to quantify other cultural characteristics. Moreover, methods to detect overlapping communities could be used to test in how much these overlap [38], and whether individuals belong to multiple communities and use different word sets in each of them [39]. There are numerous applications of our method, including social group identification, customising online experience, targeted marketing, and crowd-sourced characterisation.

Methods

Network sampling

Our sample network was formed using a process called snowball-sampling [40]: For each user sampled, all their conversational tweets (*i.e.*, tweets that are directed at another user) were recorded and any new users referenced added to a list of users from which the next user to be sampled is picked. Starting from a random user, conversational tweets, time-stamped between January 2007 to November 2009 were sampled from the Twitter web site during December 2009, yielding over 200 million messages. We ignored messages that were copies of other messages (so called retweets, which are identified by the text ‘RT’). The links in the network were bidirectional and weighted by the number of tweets sent between the two users linked.

Ranking words within a community

In order to establish which words characterise each community, we compared the fraction of users that use each word within a community with the fraction of users that used the word globally. We then assessed how unlikely it was that the difference between these two fractions could have happened by random chance. This is given by the standardised Z -score which, for each word used in community c , is

$$Z = \frac{\mu_c - \mu_g}{\sigma_g / \sqrt{N_c}},$$

where μ_c is the fraction of users in community c which have used the word, μ_g is the fraction of all users that have used it, N_c is the number of users in community c , and σ_g is

the standard deviation of usage of the word amongst all users

$$\begin{aligned}\sigma_g^2 &= \frac{1}{N}(\mu_g N(1 - \mu_g)^2 + (1 - \mu_g)N\mu_g^2) \\ &= \mu_g(1 - \mu_g),\end{aligned}$$

where N is the global number of users.

Comparing communities using a bootstrap

For each word i used by each community j we calculated its relative word usage frequency $f_j(i)$, *i.e.*, the proportion of the total word instances that were word i . Using this, we were able to measure the difference between two communities j and k by the Euclidean distance as follows,

$$d(j, k) = \sqrt{\sum_i [f_j(i) - f_k(i)]^2}.$$

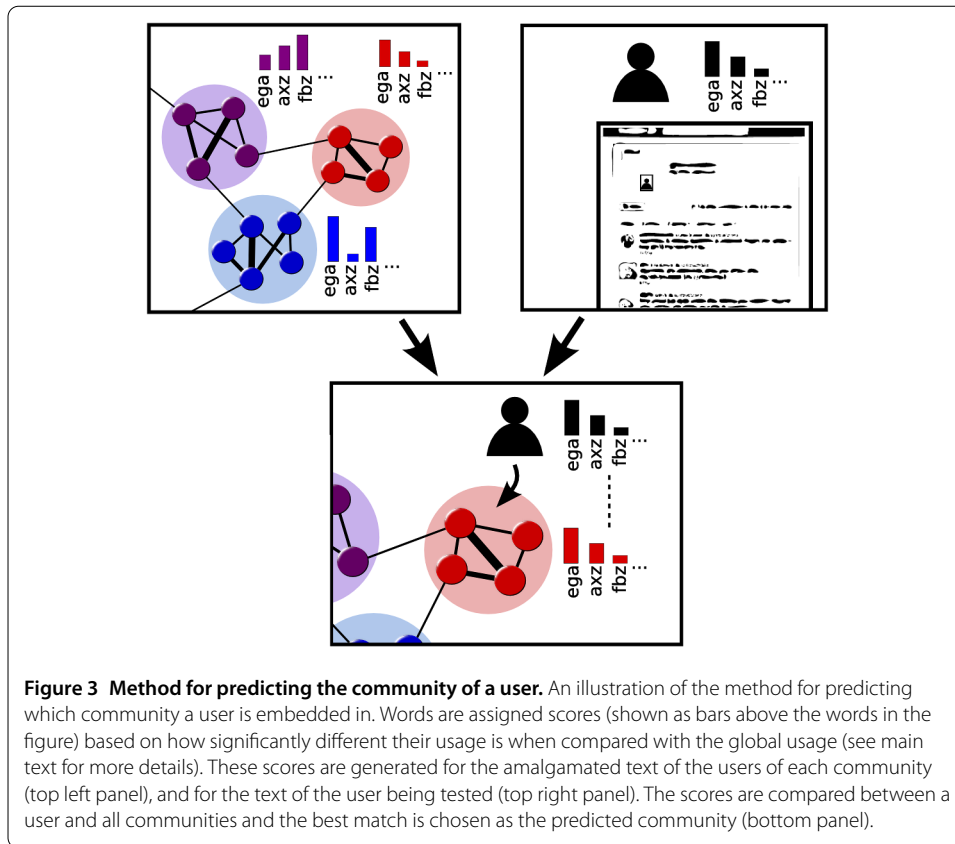
To assess the likelihood that a distance calculated could have happened by chance we performed a bootstrap. For each pair of communities we took the union of users and resampled a new pair of communities (j' and k') of the same sizes as the original pair. For each resampled pair we calculated whether the Euclidean distance of the resampled pair was greater than that of the original pair, *i.e.*, if

$$d(j', k') > d(j, k).$$

We confirmed that the distribution of resampled distances was close to a normal distribution. Over many resampled pairs, the frequency of instances when this inequality was true gave us the probability that the difference in word usage between the two communities could have happened by chance if words were randomly distributed amongst communities.

Predicting communities of individual users

To predict the communities of individual users we compared individual word usage with community word usage to select the best matching community (see Figure 3). To do this we divided the users into two equal halves: a randomly selected set of test users and a base set of the remaining users. We then randomly sampled words from test users and compared the probability that the frequency with which that user uses the sampled word would have been obtained by randomly sampling words from all users (the p -value of its relative frequency) with the same p -values in the communities of base users (*i.e.*, considering the frequency with which that word is used in the community). We then associated the community with the smallest mean difference in p -values to the test user for the sampled words with that test user. For English-speaking users, we only tested users from communities larger than 250 users.



Additional material

Additional file 1: Community word lists. The top-ranked words for English-speaking communities with more than 250 users. The communities at the highest level are numbered as in the figure, sub-communities are shown after each community. Listed are the words with *Z*-score, frequency in the whole network (*global freq*), frequency within the community (*group freq*) and the ratio of the frequency within the community to the frequency in the whole network (*ratio*).

Competing interests

Royal Holloway, University of London has filed a patent application, published as WO2012/080707, which seeks to protect many of the techniques described in this paper. Royal Holloway is currently pursuing commercial exploitation of this patent. JB and VAAJ are employees of Royal Holloway so may, in future, benefit directly or indirectly from any such exploitation.

Authors' contributions

JB conceived the study, collected the data, performed the analyses and contributed to the writing, SF led the writing, VAAJ contributed to the writing and supervised the project. All authors were involved in interpretation of the results.

Author details

¹School of Biological Sciences, Royal Holloway, University of London, Egham, TW20 0EX, UK. ²Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA. ³London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, UK.

Acknowledgements

The authors would like to express their gratitude to Edwin van Leeuwen for helpful discussions. This work was supported by the Engineering and Physical Sciences Research Council through standard research grant number EP/D002249/1, by the Biotechnology and Biological Sciences Research Council grant BB/1000151/1 (to V.A.A.J.), by the Economic and Social Research Council grant (ES/L000113/1), by the EU FP7 funded integrated project EPIWORK (grant agreement number 231807), by the US Department of Homeland Security and by the Bill and Melinda Gates Foundation.

References

1. Gumperz J (1958) Dialect differences and social stratification in a North Indian village. *Am Anthropol* 60(4):148-170
2. Labov W (1966) The linguistic variable as structural unit. *Wash Linguist Rev* 3:4-22
3. Chambers JK (1997) *Sociolinguistic theory*. Blackwell, Oxford
4. Carroll KS (2008) Puerto Rican language use on myspace.com. *Cent J* 20:96-111
5. Mæhlum B (2010) Language and social spaces. In: Auer P, Schmidt JE (eds) *Language and space: theories and methods*, vol 1. de Gruyter, Berlin
6. Labov T (1982) Social structure and peer terminology in a black adolescent gang. *Lang Soc* 11:391-411
7. Mendoza-Denton N (2007) Homegirls: language and cultural practices among Latina youth gangs. Blackwell, Oxford
8. Milroy L (1980) *Language and social networks*. Blackwell, Oxford
9. Porter MA (2009) Communities in networks. *Not Am Math Soc* 56(9):1164-1166
10. Fortunato S (2010) Community detection in graphs. *Phys Rep* 486:75-174
11. Newman MEJ (2006) Modularity and community structure in networks. *Proc Natl Acad Sci USA* 103(23):8577-8582
12. Traud AL, Kelsic ED, Mucha PJ, Porter MA (2011) Comparing community structure to characteristics in online collegiate social networks. *SIAM Rev* 53(3):526-543
13. Bryden J, Funk S, Geard N, Bullock S, Jansen VAA (2011) Stability in flux: community structure in dynamic networks. *J R Soc Interface* 8(60):1031-1040
14. Kucera H, Francis WN (1982) *Frequency analysis of English usage: lexicon and grammar*. Houghton Mifflin, Boston
15. Michel J, Shen YK, Aiden AP, Veres A, Gray MK, The Google Books Team, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Pinker S, Nowak MA, Aiden AL (2010) Quantitative analysis of culture using millions of digitized books. *Science* 331:176-182
16. Kumpula JM, Onnela JP, Saramaki J, Kaski K, Kertesz J (2007) Emergence of communities in weighted networks. *Phys Rev Lett* 99:228701
17. Blondel VD, Guillaume J, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *J Stat Mech Theory Exp* 2008(10):P10008
18. Reichardt J, Bornholdt S (2007) Partitioning and modularity of graphs with arbitrary degree distribution. *Phys Rev E* 76:015102
19. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci USA* 105(4):1118-1123
20. Rosvall M, Axelsson D, Bergstrom CT (2009) The map equation. *Eur Phys J Spec Top* 178:13-23
21. Rosvall M, Bergstrom CT (2011) Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLoS ONE* 6(4):e18209. <http://dx.doi.org/10.1371/journal.pone.0018209>
22. Lancichinetti A, Fortunato S (2009) Community detection algorithms: a comparative analysis. *Phys Rev E* 80:056117
23. Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. *Inf Process Manag* 24(5):513-523. <http://www.sciencedirect.com/science/article/pii/0306457388900210>
24. Heckathorn DD (1997) Respondent-driven sampling: a new approach to the study of hidden populations. *Soc Probl* 44:174-199
25. Kemeny JG, Snell JL (1960) *Finite Markov chains*. Van Nostrand, Princeton
26. Sunnafrank M, Ramirez A Jr (2004) At first sight: persistent relational effects of get-acquainted conversations. *J Soc Pers Relatsh* 21(3):361-379
27. McElreath R, Boyd R, Richerson PJ (2003) Shared norms and the evolution of ethnic markers. *Curr Anthropol* 44:122-129
28. McPherson JM, Smith-Lovin L, Cook J (2001) Birds of a feather: homophily in social networks. *Annu Rev Sociol* 27:415-444
29. Fehr E, Fischbacher U (2004) Social norms and human cooperation. *Trends Cogn Sci* 8(4):185-190
30. Centola D, Gonzalez-Avella JC, Eguiluz VM, Miguel MS (2007) Homophily, cultural drift, and the co-evolution of cultural groups. *J Confl Resolut* 51(6):905-929
31. Boyd R, Richerson PJ (2009) Culture and the evolution of human cooperation. *Philos Trans R Soc Lond B* 364(1533):3281-3288
32. Hill KR, Walker RS, Božičević M, Eder J, Headland T, Hewlett B, Hurtado AM, Marlowe F, Wiessner P, Wood B (2011) Co-residence patterns in hunter-gatherer societies show unique human social structure. *Science* 331(6022):1286-1289
33. Funk S, Gilad E, Watkins C, Jansen VAA (2009) The spread of awareness and its impact on epidemic outbreaks. *Proc Natl Acad Sci USA* 106(16):6872-6877
34. Salathé M, Jones JH (2010) Dynamics and control of diseases in networks with community structure. *PLoS Comput Biol* 6(4):e1000736
35. Salathé M, Bonhoeffer S (2008) The effect of opinion clustering on disease outbreaks. *J R Soc Interface* 5(29):1505-1508
36. Efferson C, Lalive R, Fehr E (2008) The coevolution of cultural groups and ingroup favoritism. *Science* 321(5897):1844-1849
37. Lazer D, Pentland A, Adamic L, Aral S, Barabasi A, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Alstyn MV (2009) Computational social science. *Science* 323(5915):721-723
38. Ahn YY, Bagrow JP, Lehmann S (2010) Link communities reveal multiscale complexity in networks. *Nature* 466:761-764
39. Clark HH, Brennan SE (1991) Grounding in communication. In: Resnick LB, Levine JM, Teasley SD (eds) *Perspectives on socially shared cognition*. Am Psychol Assoc, Washington
40. Goodman LA (1961) Snowball sampling. *Ann Math Stat* 32:148-170

doi:10.1140/epjds15

Cite this article as: Bryden et al.: Word usage mirrors community structure in the online social network Twitter. *EPJ Data Science* 2013 2:3.