Engagement and Interactions of Language Communities on Twitter

Nabeel Albishry
Department of Computer Science
University of Bristol
Bristol, UK
Email: n.albishry@bristol.ac.uk

Theo Tryfonas
Department of Civil Engineering
University of Bristol
Bristol, UK
Email: theo.tryfonas@bristol.ac.uk

Tom Crick
Department of Computing
Cardiff Metropolitan University
Cardiff, UK
Email: tcrick@cardiffmet.ac.uk

***Abstract*—Whilst emerging research is providing insight into the factors that promote the propagation of information in online social networks following significant events, such as rioting and terrorism, this paper evaluates the extent to which different language communities engage and interact. We present our analysis of online interactions in various languages that took place on the social networking site Twitter during the Baltimore protests in April 2015 in the USA.**

By utilising language information from user profiles (*N*=716,494) and status updates (*N*=1,257,065) relating to the Baltimore protests to identify and categorise communities, we are able to provide insight into the pattern of their interactions, as well as constructing their network graphs. The results show that the nature of the event is reflected on the engagement degree and wider interaction of communities. This analysis of language communities may also help in deciding which group of users to engage with, and hence increase the chance of influential action when participating on Twitter conversations.

I. Introduction

A. Online Social Networks

In recent years, online social networks (OSNs) have been utilised as means to express ideas and opinions, spread information about events, or even stimulate and propagate calls for civic engagement and societal action. Social networking sites such as Twitter, Facebook, LinkedIn and YouTube have also empowered individuals to promote their viewpoints and interests – professional or otherwise – to a broad and diverse global audience. The engagement of certain demographics with social networks offers the opportunity for researchers interested in observing and interpreting society to apply established theory and methods to an emerging digital culture.

To satisfy the demand for various types of communities, interactions and engagement, there are now vast numbers of social media sites and platforms[1], along with a number of attempted categorisations. By 2018, there will be an estimated 2.5 billion active social network users (up from 1.9 billion in 2014); they are producing massive amounts of data (volume) on a real-time basis (velocity) with implicit sociological attributes such as beliefs, opinions, sentiments, behaviours, structures and influences (variety) [1]. These data exhibit the key traits of what is now referred to as big data: volume, velocity and variety [2]. In this age of big data and an increasingly interconnected digital society, there is a new challenge – the application of robust and scalable methods and tools that can be applied to digitised social behaviour generated via social networks so as to be able to efficiently analyse big social data to provide insight into real-world events and actions [1], [3].

Recent work [4]–[8] has analysed what people say on social media to identify distinctive words, phrases, and topics as functions of known attributes of people such as gender, age, location, or psychological characteristics. This can thus be collated and aggregated, inferring gender, age, location and sentiments, from social media data. Potential negative implications of these approaches include the fact that they can be easily applied to large numbers of people or groups in society without obtaining their explicit consent or even being aware it is being done. Data-driven commercial companies, governmental entities, or even one's followers or friends are able to use software to infer personality and other attributes – such as sexual orientation or political affiliations – that an individual may have decided not to share [2], [9].

There are various projects that have used Twitter corpora and related datasets to make predictions about elections [10], stock markets [11], and crimes and policing [12]–[14]. Twitter played an important role during what was then known as the "Arab Spring", which has been extensively examined in the social network analysis domain [15]–[19]. While the use of Twitter data has been demonstrated to provide insight – and sociologically relevant demographics [20] – into major social and physical events such as riots [21] and terror attacks [22], often all is not what it may seem; for instance many tweets may not a crowd make [23].

B. Languages and Communities

Despite the widespread engagement with Twitter globally, little research has investigated the differences amongst users of various languages; there is a tendency to assume that the behaviours of English users generalise to other language users [24]. Language has featured as a facet of research on the geographies of Twitter networks [25], especially whether offline geography still matter in online social networks [26]. Linguistic-inspired studies have been done on hashtags [27], as

[1]This list is by no means exhaustive: http://en.wikipedia.org/wiki/Listofsocialnetworkingwebsites

well as the volume and proportional of tweets in English and Arabic, as part of an analysis of the Arab Spring [19]. Nevertheless, language is clearly a vital component of affiliation and discourse on the web [28], with the creation and curation of emerging multi-lingual networks and communities, representing well-established creative, cultural and socio-economic norms, including for minority languages such as Welsh [29].

C. Social Network Analysis

In the social network analysis (SNA) domain, centrality measures provide the ability to assess network graphs that are constructed from collected data (for example, tweets). Selection of these centrality measures is dependent on the goal of the analysis; for example, the degree of node helps to identify nodes with high number of connections within the network [30]–[32]. In a representation of a real world network, this metric may help to identify highly connected persons, such as political leaders, sports stars or celebrities, who are potential "information spreaders" [33]–[35]. Centrality measures such as degrees, betweenness, clustering coefficient, modularity and cliques have been used in many projects to measure influence or detect the emergence of new communities [14], [36].

Clustering users in communities has been an important analytic factor in social networking analysis; numerous work has focused on clustering users based on their locations. However, for the sake of anonymity, many users tend not to disclose information about their identity, such as locations [37]. It has also been reported in the literature that geotagged tweets are generally low in number [38]–[40], the exponential growth in social media over the past decade has been joined by the rise of location as a central organising theme [23] of how users engage with online information services and, more importantly, with each other [41], [42].

D. Users and Location

It is important to understand how geotagging works in Twitter. The 'place' entity included in a Twitter status does not necessarily indicate precisely where the actual posting was made, as stated in the Twitter API documentation[2]:

> "*Tweets associated with places are not necessarily issued from that location but could also potentially be about that location.*

For the sake of anonymity many users tend not to disclose information about their identity, particularly locations; this has also been supported by the literature that geotagged tweets are generally low in number [37]. An alternative location-based option to consider is based on profile location, but this may not serve the need for location clustering for a multitude of reasons, especially with a significant proportion of Twitter users not setting their profile location [43] (discussed in more detail in Section III-A).

[2]https://dev.twitter.com/overview/api/places

E. Overview of Paper

The techniques we introduce in this paper are based on language settings in users' profiles and those for statuses[3]. The remainder of this paper is organised as follows: Section II introduces the context of the Baltimore protest case study and the development of events in alignment with activity on Twitter. In Section III we present the techniques we have used to identify and analyse language communities and networks, along with the results. Finally, Section IV concludes the paper with a wider discussions and a summary of the potential application of our approach.

II. Context and Events Timeline

Following the peaceful funeral of Freddie Gray that took place on the morning of Monday 27 April 2015 in Baltimore, Maryland, USA, a protest hit the city. According to the timeline published on the CNN website "*The city exploded on Monday after the funeral of Freddie Gray, a 25-year-old black man who mysteriously died on April 19, a week after Baltimore Police arrested him.*" [44]. The nature of the Baltimore protests is a good representation of a planned event in which a sudden escalation of violence hits a geographical area. The event manifested itself on Twitter as `#BaltimoreRiots`, and resulted in more than 1,250,000 status updates.

Figure 1 and Figure 2 present how the event manifested itself on Twitter once a "purge" was scheduled. We can see that what was happening on the ground was quickly reflected on the activity in Twitter, as Figure 2 indicates. More detailed analysis reveals that within one hour the topic started to go "viral"; more precisely, at approximately 15:00 at which the "purge" was scheduled. The topic jumped from roughly 1,200 to 8,000 tweets per hour. Then, it peaked with 98,000 between 22:00 and 23:00.

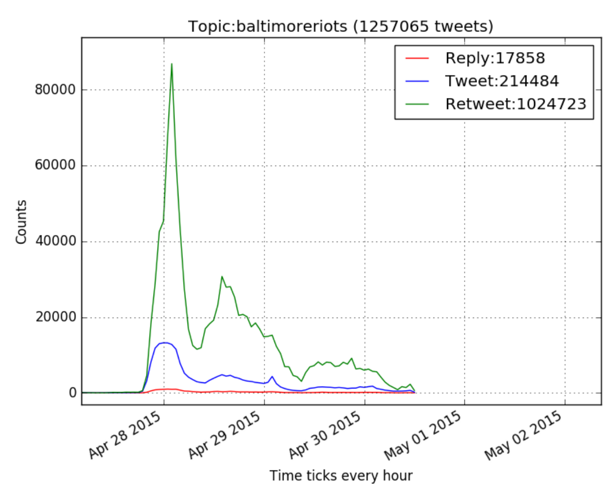

Figure 1. Overall activity for `#BaltimoreRiots`.

[3]The term 'status' is a generic term used to refer to any Twitter post (tweet, retweet, reply, or quote).

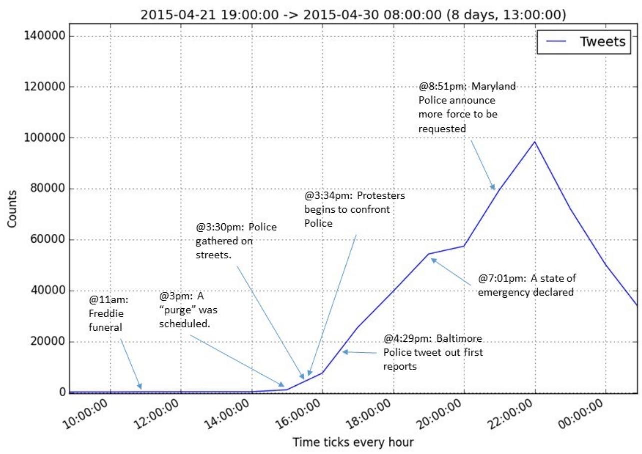

Figure 2. Main events during the lifetime of `#BaltimoreRiots`.

III. Language Communities

Analysis of language communities begins with two basic techniques. The first is to classify statuses based on their languages. The status language is extracted from the '*lang*' entity inside status objects. Language used in posting defines which community the status was meant for; a tweet written in Turkish, for example, is meant for the Turkish-speaking community. Output from this will be referred to as 'posting communities'. The second analysis is to classify users into different communities based on their profile languages, regardless of the posting language they used. Output from this technique will be referred to as 'profile communities'. As we will see in the following sections, a posting community does not necessarily indicate the profile community for a user. Therefore, the second step is to examine the relationship between profile and posting communities. We will also explore relationships amongst profile communities, in term of action-reaction. We will investigate the intra- and inter-profile communities interactions by constructing network graphs and generate a visual representation.

A. Locations

As mentioned in Section I, for the sake of anonymity many users tend not to disclose information about their identity, particularly locations; this has also been supported by the literature that geotagged tweets are generally low in number [37]. We took the step to verify this claim in our datasets; in the best cases, the ratio of geotagged tweets did not exceed 2%. In the case of the `#BaltimoreRiots` dataset, only 1% of collected statuses were associated with places. Moreover, out of this geotagged subset, only 4% were associated with the city where the event took place (Baltimore).

An alternative location-based option to consider is based on profile location, but it still does not serve the need for location clustering for a multitude of reasons. Firstly, we found that less than 45% of users have set their profile location, which is in line with other studies [43]. Secondly, although Twitter suggests certain presets for setting profile location, users are given the option to enter any text they wish; this results in a considerable amount of noise.

B. Posting Communities

In the `#BaltimoreRiots` case, there were 38 posting languages. As we can see in Figure 3, English was the dominating language by far. Interestingly, results also show that language of more than 41,000 (3%) statuses could not be identified. When investigated, those statuses mostly do not contain text other than hashtags, pictures or URLs. Although, this is not a big portion, it came second after English. Although this category shows an interesting case in which qualitative content analysis would be involved, it is beyond this study and will not be covered here.

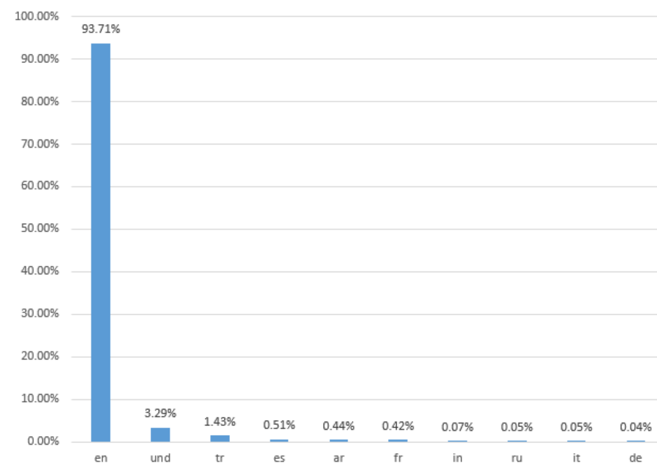

Figure 3. Most frequently used languages in `#BaltimoreRiots`: (*en:* English; *es:* Spanish; *tr:* Turkish; *fr:* French; *en-gb:* British English; *ar:* Arabic; *de:* German; *ru:* Russian; *it:* Italian; *pt:* Portuguese)

C. Users' Language Communities

In the majority of cases, users choose to pick a language for their Twitter profile settings. In our dataset we found that out of 716,494 users, only 45 had not chosen any language. However, the language entity returned by the API for those cases is the initial placeholder text "*Select Language...*" or a translated version that might provide hints regarding the user language community. Figure 4 shows that about 94% of the users came from '*en*' profile community.

From these two outputs, we can see that nearly all of the topic activity came from one particular community using one particular language. This extreme pattern may accompany extreme and geographically constrained real world events such as riots and terror attacks.

D. Profile-Posting Graph

To investigate whether the '*en*' posting community is linked to particular profile communities, we constructed a bipartite graph as presented in Figure 5, representing the profile-posting language network. In this graph, nodes that are prefixed by

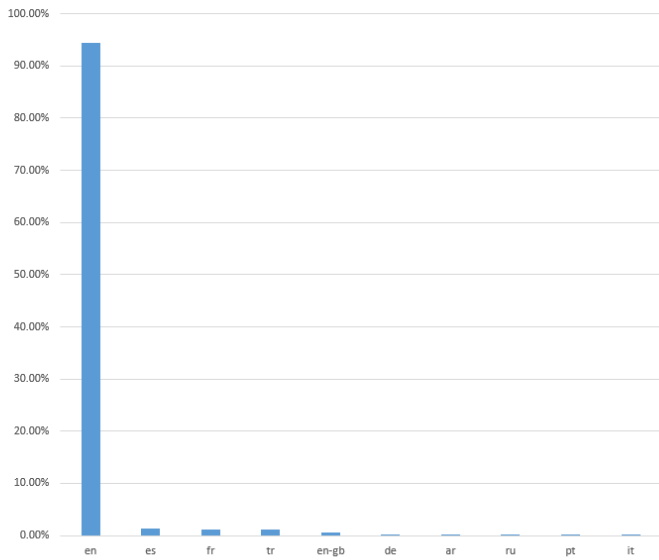

Figure 4. Top 10 most active communities in `#BaltimoreRiots`

"*p_*" represent profile language community, and nodes that are prefixed by "*s_*" represent status language community. Size of node represents the weighted degree, whereas colour represents the outdegree; the darker the colour, the higher outdegree. The graph confirms the domination pattern we highlighted earlier; furthermore, it shows the relationships between the profile and posting communities.

For an extreme case of one dominating posting language, we wanted to investigate participation of different communities. We thus filtered out all non-'*en*' statuses, and then identified different profile communities with the resultant set. For each community, we classified statuses into two sets: *actions* and *reactions*; this result is shown in Table I. This shows the highest scoring communities, where the first column represents the category of status (action or reaction), community column represents profile language community, and last one shows percentage of '*en*' posts by that community.

Category	**Community**	**%**
Reaction	*en*	81.08
Action	*en*	15.43
Reaction	*es*	0.68
Reaction	*fr*	0.59
Reaction	*en-gb*	0.47
Reaction	*tr*	0.19
Reaction	*de*	0.18
Reaction	*pt*	0.13

Table I
ACTIVITY AND CATEGORIES OF MOST ACTIVE PROFILE LANGUAGE COMMUNITIES

From the results above we can infer that there is a dominating player in both domains: posting languages and profile communities. Therefore, for the case of `#BaltimoreRiots`, we can conclude that the case was substantially localised.

E. Temporal Communities Activity

We wished to explore communities' activity over time, apart from the overall activity. Figure 7 represents a heat map for posts per hour for each profile community over the lifetime of `#BaltimoreRiots`. This sort of mapping would help in identifying times at which communities are active.

When refined with posting language, this technique would be useful in identifying when to engage into conversations, which language community to target, and by which language. For example, presuming that we want to participate in a trending topic hashtag that we are interested in; this analysis technique could help us make our post more direct and focused. By finding which language is mostly used in posting, we will be able to know in which language the tweet would be more effective. Also, we might want to direct the message to certain language community, be that to influence a very active or re-activate a quiet one.

F. Reaction Networks

Another important perspective to identify and capture is how different profile communities relate to each other in terms of reactions. Therefore, we constructed the graph in Figure 6 to show interactions amongst language communities. Edges are directed and are drawn from reacting nodes (retweeter, or replier) to the original acting node (tweeter). As we are not interested in reactions from community to itself, we eliminated self-loop edges from the graph. Node size represents weighted indegree; we have also generated the degree measures for the various language communities, as shown in Figure 8.

IV. CONCLUSIONS

This paper presented a study in identifying languages used, language communities and their engagement and interactions on the Twitter platform with respect to real world events, in this instance using the Baltimore protests in the USA in April 2015. As we discussed in Section III, the nature of the event (e.g. being a local or global) may be reflected on community conversations on Twitter. We found that most of posting activity comes from the main community (the language community in which the incident has happened or tightly related to). This is especially the case when the online conversations are triggered by a real world incident. The same for posting languages, users mostly to use the language of the main community. Although most of topic activity comes from that main community, we noticed that, with local events, other communities work, mostly, as information spreaders.

We also presented a network graph (using Gephi[4] and the Networkx Python package[5]) showing how language communities relate to each other in the form of action-reaction (action: tweet, reaction: retweet, reply, quote). Another interesting graph that we produced to show relations between profile and posting communities. We find that this graph is important to facilitate comparing users defined profile language with their

[4] https://gephi.org/
[5] https://networkx.github.io/

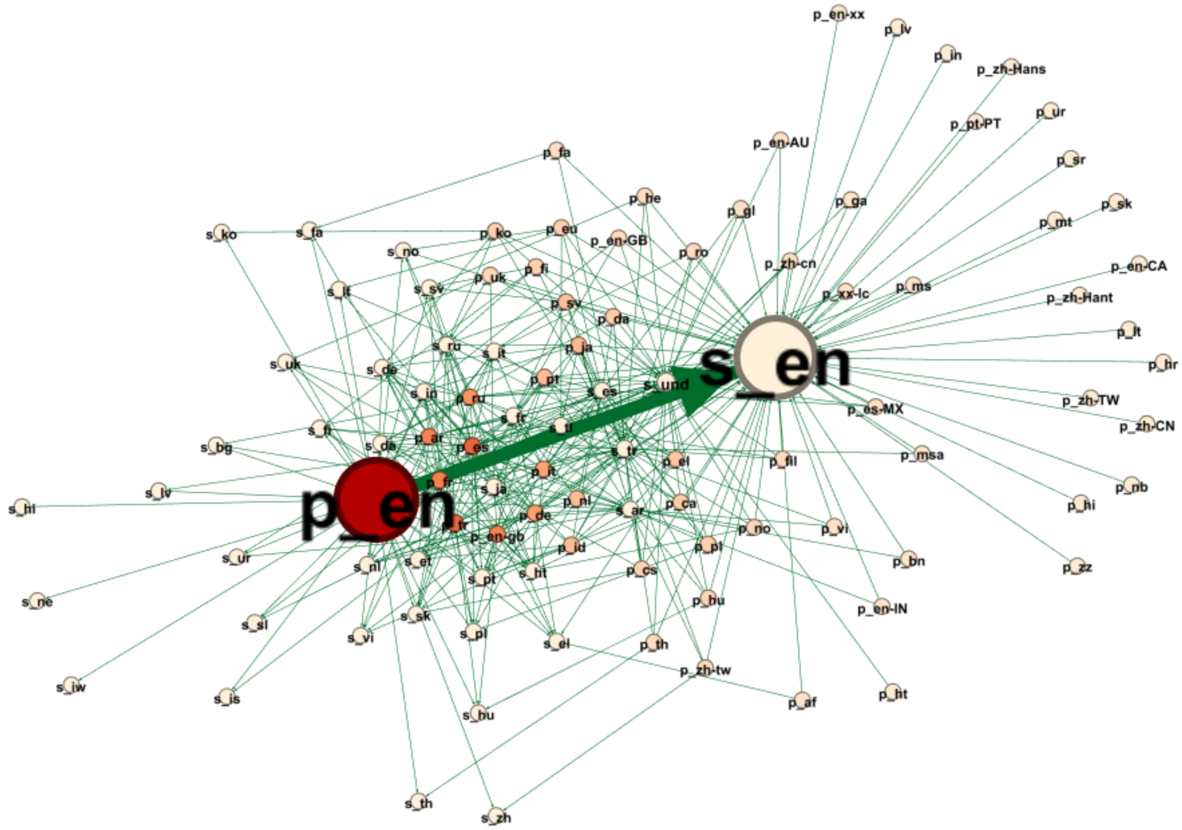

Figure 5. Profile-posting network graph

posting language. Some event might be termed as ‘partially scheduled’ as their end was different to how they were planned in the first place. In such tense situation, we noticed that diversity of languages and communities are very low, and there always be a dominating community and language.

The method we presented here can be used in identifying how communities interact with one another, which ones are most active, which languages are mostly used, and at what time. Applying these techniques on data pouring from the Twitter Stream API[6] would be applicable to a wide number of domains. For example, these methods can be used in social network marketing and publicity to increase the probability of influential posts. In practice, for a given `#<Brand>`, by monitoring the activity of different language community, one can decide the time to post well-tailored tweets targeting certain communities. This can be fine-tuned further by mentioning key players in that community, e.g. users with high closeness scores.

[6]https://dev.twitter.com/streaming/overview

Moreover, within certain contexts, the order of applying these two classifications (posting and profile) will generate different results. For example, taking one profile community and dividing it into different posting communities shows the number of languages this community may use, and hence degree of openness and reachability. A possible scenario for governments, politicians or campaigners would be to use this method to measure to what extent other languages are used within a profile community. It may also show how users associate themselves with one community in their profile while using other languages. Monitoring unusual activity for secondary languages may help to uncover important messages or opinions that could not be openly expressed, for a variety of reasons, to the rest of the profile community.

For the social network analysis domain, this method provides a different perspective for influence analysis. Endorsement from different profile communities cannot be measured similar to those coming from the same community. For example, in a controversial Arabic topic, we noticed that high support came from other profile communities.

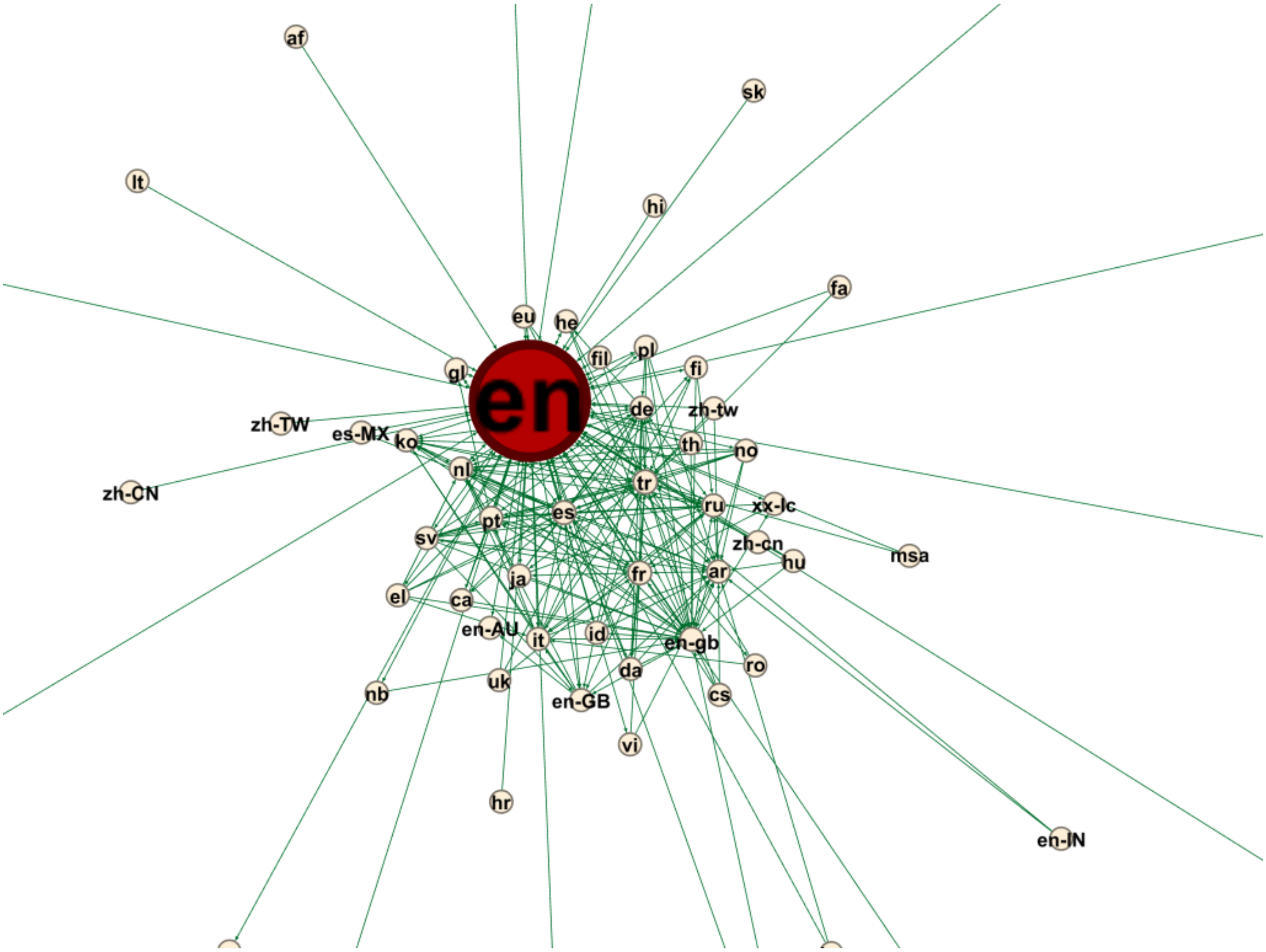

Figure 6. Profile-profile network graph

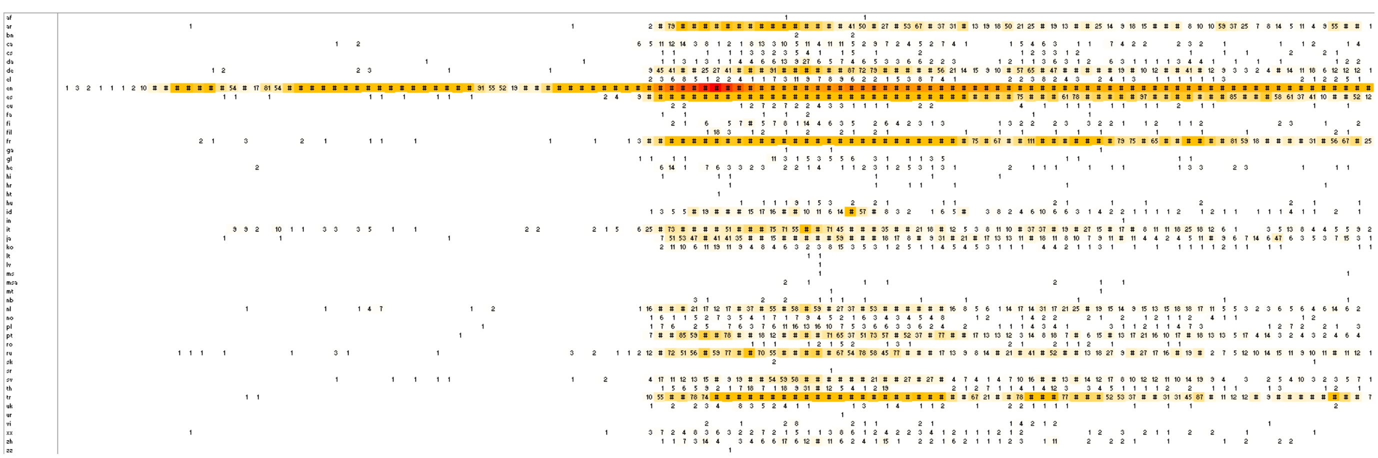

Figure 7. Temporal activity of profile communities

For future work, we plan to add further classifications to the reaction network (as presented in Section III-F). We believe that differentiation between endorsements (e.g. retweets) and other reactions may provide further insight into the networks and communities. Furthermore, we will apply the methods presented in this paper on other high-profile event/discussion

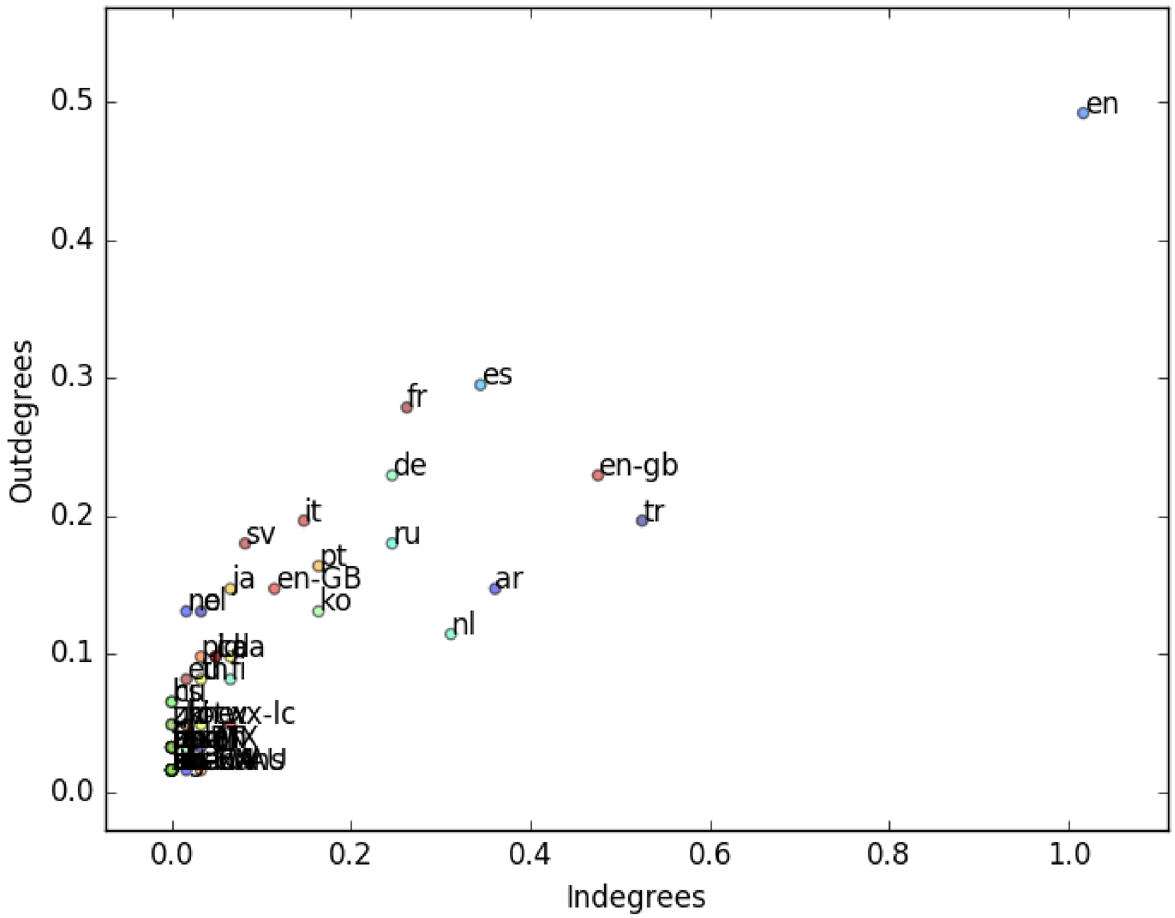

Figure 8. Degree measures for the various language communities

datasets in different domains or contexts, such as for sports, music contests and humanitarian actions.

ACKNOWLEDGMENT

This work has been supported by a doctoral research scholarship for Nabeel Albishry from King Abdulaziz University, Kingdom of Saudi Arabia.

REFERENCES

[1] P. Burnap, O. Rana, M. Williams, W. Housley, A. Edwards, J. Morgan, L. Sloan, and J. Conejero, "COSMOS: Towards an integrated and scalable service for analysing social media on demand," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 30, no. 2, pp. 80–100, 2015.

[2] UK Parliamentary Office of Science and Technology, "Social Media and Big Data," Tech. Rep. Report 460, March 2014.

[3] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne, "Computational Social Science," *Science*, vol. 323, no. 5915, pp. 721–723, 2009.

[4] B. Blamey, T. Crick, and G. Oatley, "R U :-) or :-(? Character- vs. Word-Gram Feature Selection for Sentiment Classification of OSN Corpora," in *Research and Development in Intelligent Systems XXIX*. Springer, 2012, pp. 207–212.

[5] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach," *PLoS ONE*, vol. 8, no. 9, 2013.

[6] B. Blamey, T. Crick, and G. Oatley, "'The First Day of Summer': Parsing Temporal Expressions with Distributed Semantics," in *Research and Development in Intelligent Systems XXX*. Springer, 2013, pp. 389–402.

[7] G. Oatley and T. Crick, "Changing Faces: Identifying Complex Behavioural Profiles," in *Human Aspects of Information Security, Privacy and Trust*, ser. Lecture Notes in Computer Science. Springer, 2014, vol. 8533, pp. 282–293.

[8] G. Oatley, T. Crick, and M. Mostafa, "Digital Footprints: Envisaging and Analysing Online Behaviour," in *Proceedings of 2015 Symposium on Social Aspects of Cognition and Computing Symposium (SSAISB)*, 2015.

[9] R. Lambiotte and M. Kosinski, "Tracking the Digital Footprints of Personality," *Proceedings of the IEEE*, vol. 102, no. 12, pp. 1934–1939, 2014.

[10] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," in *Proceedings of the 4th International AAAI Conference on Web and Social Media (ICWSM)*, 2010.

[11] X. Zhang, H. Fuehres, and P. A. Gloor, "Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear"," *Procedia – Social and Behavioral Sciences*, vol. 26, pp. 55–62, 2011.

[12] M. S. Gerber, "Predicting crime using Twitter and kernel density estimation," *Decision Support Systems*, vol. 61, pp. 115–125, 2014.

[13] G. Oatley and T. Crick, "Exploring UK Crime Networks," in *Proceedings of International Symposium on Foundations of Open Source Intelligence and Security Informatics (FOSINT-SI)*, 2014.

[14] ——, "Measuring UK Crime Gangs: A Social Network Problem," *Social Network Analysis and Mining*, vol. 5, no. 1, 2015.

[15] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce, and D. Boyd, "The Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions," *International Journal of Communication*, vol. 5, pp. 1375–1405, 2011.

[16] P. N. Howard, A. Duffy, D. Freelon, M. M. Hussain, W. Mari, and M. Maziad, "Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring?" 2011, available at: http://dx.doi.org/10.2139/ssrn.2595096.

[17] F. Comunello and G. Anzera, "Will the revolution be tweeted? A conceptual framework for understanding the social media and the Arab Spring," *Islam and Christian-Muslim Relations*, vol. 23, no. 4, pp. 453–470, 2012.

[18] G. Wolfsfeld, E. Segev, and T. Sheafer, "Social Media and the Arab Spring: Politics Comes First," *The International Journal of Press/Politics*, vol. 18, no. 2, pp. 115–137, 2013.

[19] A. Bruns, T. Highfield, and J. Burgess, "The Arab Spring and Social Media Audiences: English and Arabic Twitter Users and Their Networks," *American Behavioral Scientist*, vol. 57, no. 7, pp. 871–898, 2013.

[20] L. Sloan, J. Morgan, W. Housley, M. L. Williams, A. Edwards, P. Burnap, and O. F. Rana, "Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter," *Sociological Research Online*, vol. 18, no. 3, 2013.

[21] R. Procter, J. Crump, S. Karstedt, A. Voss, and M. Cantijoch, "Reading the riots: what were the police doing on Twitter?" *Policing and Society*, vol. 23, no. 4, pp. 413–436, 2013.

[22] P. Burnap, M. L. Williams, L. Sloan, O. F. Rana, W. Housley, A. Edwards, V. Knight, R. Procter, and A. Voss, "Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack," *Social Network Analysis and Mining*, vol. 4, no. 1, 2014.

[23] Y. Liang, J. Caverlee, Z. Cheng, and K. Y. Kamath, "How big is the crowd?: event and location based population modeling in social media," in *Proceedings of the 24th ACM Conference on Hypertext and Social Media (HT'13)*, 2013, pp. 99–108.

[24] L. Hong, G. Convertino, and E. H. Chi, "Language Matters In Twitter: A Large Scale Study," in *Proceedings of the 5th International AAAI Conference on Web and Social Media (ICWSM)*, 2011.

[25] Y. Takhteyev, A. Gruzd, and B. Wellman, "Geography of Twitter networks," *Social Networks*, vol. 34, no. 1, pp. 73–81, 2012.

[26] J. Kulshrestha, F. Kooti, A. Nikravesh, and K. P. Gummadi, "Geographic Dissection of the Twitter Network," in *Proceedings of the 6th International AAAI Conference on Web and Social Media (ICWSM)*, 2012.

[27] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. Gonçalves, and F. Benevenuto, "Analyzing the dynamic evolution of hashtags on Twitter: a language-based approach," in *Proceedings of the Workshop on Languages in Social Media (LM'11)*, 2011, pp. 58–65.

[28] M. Zappavigna and J. R. Martin, *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*. Continnuum, 2012.

[29] E. Gruffydd Jones and E. Uribe-Jongbloed, Eds., *Social Media and Minority Languages: Convergence and the Creative Industries*. Multilingual Matters Ltd, 2013.

[30] S. P. Borgatti and M. G. Everett, "Models of core/periphery structures," *Social Networks*, vol. 21, no. 4, pp. 375–395, 2000.

[31] M. Rombach, M. A. Porter, J. H. Fowler, and P. J. Mucha, "Core-Periphery Structure in Networks," *SIAM Journal on Applied Mathematics*, vol. 74, no. 1, pp. 167–190, 2014.

[32] W. Liu, M. Pellegrini, and X. Wang, "Detecting Communities Based on Network Topology," *Scientific Reports*, vol. 4, no. 5739, 2014.

[33] M. Cha, F. Benevenuto, H. Haddadi, and K. Gummadi, "The World of Connections and Information Flow in Twitter," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 42, no. 4, pp. 991–998, 2012.

[34] J. Borge-Holthoefer, A. Rivero, and Y. Moreno, "Locating privileged spreaders on an online social network," *Physical Review E*, vol. 85, no. 066123, 2012.

[35] J.-X. Zhang, D.-B. Chen, Q. Dong, and Z.-D. Zhao, "Identifying a set of influential spreaders in complex networks," *Scientific Reports*, vol. 6, no. 27823, 2016.

[36] A. Willis, A. Fisher, and I. Lvov, "Mapping networks of influence: tracking Twitter conversations through time and space," *Participations: Journal of Audience & Reception Studies*, vol. 12, no. 1, pp. 494–530, 2015.

[37] R. Kang, S. Brown, and S. Kiesler, "Why do people seek anonymity on the internet?: informing policy and design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 2657–2666.

[38] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the Sample Good Enough? Comparing Data from Twitters Streaming API with Twitters Firehose," in *Proceedings of the 7th International AAAI Conference on Web and Social Media (ICWSM)*, 2013, pp. 400–408.

[39] L. Tan, S. Ponnam, P. Gillham, B. Edwards, and E. Johnson, "Analyzing the impact of social media on social movements: A computational study on Twitter and the Occupy Wall Street movement," in *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2013.

[40] S. Kumar, F. Morstatter, and H. Liu, *Twitter Data Analytics*. Springer, 2014.

[41] Z. Cheng, J. Caverlee, and K. Lee, "You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users," in *Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM'10)*. ACM Press, 2010, pp. 759–768.

[42] J. Caverlee, Z. Cheng, D. Z. Sui, and K. Yeswanth Kamath, "Towards Geo-Social Intelligence: Mining, Analyzing, and Leveraging Geospatial Footprints in Social Media," *IEEE Data Engineering Bulletin*, vol. 36, no. 3, pp. 33–41, 2013.

[43] M. Graham, S. A. Hale, and D. Gaffney, "Where in the World Are You? Geolocation and Language Identification in Twitter," *The Professional Geographer*, vol. 66, no. 4, pp. 568–578, 2014.

[44] Wikipedia, "2015 Baltimore protests," https://en.wikipedia.org/wiki/2015_Baltimore_protests, 2015, (accessed 2016-08-14).