Sociolect-Based Community Detection

William N. Reynolds[a d], William J. Salter[b], Robert M. Farber[c], Courtney Corley[e], Chase P. Dowling[e],
William O. Beeman[f], Lynn Smith-Lovin[g] and Joon Nak Choi[h]

[a] Least Squares Software, Inc. 12231 Academy Rd. NE 301-192, Albuquerque, NM 8711 bill@leastsquares.com
[b] Strategic Solutions Consulting, 3 Elm Street Harvard, MA 01451
[c] BlackDog Endeavors, LLC, 3004 Shawnee Dr NW, Gig Harbor, WA 98355
[d] Arctan, Inc.2200 Wilson Blvd, Suite 102-150 Arlington, VA 22201
[e] Department of Energy, Pacific Northwest National Laboratories, Richland, WA 99352
f Department of Anthropology, University of Minnesota, Minneapolis, MN 55455
[g]Department of Sociology, Duke University, 348A Soc/Psych Building Campus Box 90088
[h]Department of Management, The Hong Kong University of Science and Technology, 九龍清水灣 Hong Kong

Abstract—**"Sociolects" are specialized vocabularies used by social subgroups defined by common interests or origins. We applied methods to retrieve large quantities of Twitter data based on expert-identified sociolects and then applied and developed network-analysis methods to relate sociolect use to network (sub-) structure. We show that novel methods including consideration of node populations, as well as edge counts, provide substantially enhanced performance compared to standard assortativity. We explain these methods, show their utility in analyzing large *corpora* of social media data, and discuss their further extensions and potential applications.**

Keywords—sociolect; community detection; network analysis; assortativity; social media analysis

I. Introduction

Over the past decade, there has been tremendous interest in quantitatively analyzing the rapidly growing phenomenon of social media communications. Potential application areas include marketing, intelligence analysis and information operations, and public health tracking and education campaigns. The foundational work in social networks comes from the social sciences, decades before the existence of electronic social media. See, for example, [1] [2]. The quantitative network science community (largely physicists) has developed a range of methods, beginning with the ER model of random graphs [3], extended by identification of the properties of small-world [4] and scale-free [5] networks, and by a proliferation of empirical and theoretical studies across a range of disciplines.

The basic goal of the research reported here is to develop novel methods that combine social science theories and insights with powerful network analysis techniques. Such methods can be of both theoretical interest and practical utility in a variety of domains, including intelligence and military applications. Our approach is to operationalize social science theories to apply to massive amounts of social media data and then to combine network analysis methods with those operationalizations. This paper summarizes work performed during the first eight months of the project, focusing on methodological innovations.

In particular, we investigate how patterns of language use in a social network are related to link structure, thus combining *content* information – language use – with *structural* information. To our surprise, we have not been able to find other research that explicitly explores this relationship.

Sociolects are specialized vocabularies used by professional groups, by hobbyists (e.g., knitters, bow hunters, comic collectors), and by other groups that share an interest in a given topic (e.g., "gamers," Grateful Dead fans, Civil War reenactors) [6] [7]. As defined by Louwerse [8], sociolects are "similarities in the language use of a group of individuals."

We operationalized the social science concept of sociolects [6] for analysis of groups in social networks [9] based on Twitter [10] data. Our hypothesis is that individuals using a sociolect will tend to be more closely linked than individuals who do not. This is an example of *homophily*, [11] informally, the idea that "birds of a feather flock together;" in this case the "feathers" are group-specific patterns of language usage.

Likelihood of linkage in networks or graphs is a very active research area, focusing on identifying "communities" (or "modules" or "clusters"). We briefly review some key methodological issues in community detection to set the stage for our analysis. Conceptually, communities are defined as sets of nodes more densely connected internally than with the rest of the graph, but this seemingly simple idea has proven both conceptually and computationally challenging. (See [12] for an extensive review, also [13].) In the work reported here, we addressed a simpler version of the problem: rather than attempting to decompose the networks into communities, we are concerned with assessing the link-based "communityness" of sets of nodes based on their sociolect use. In future work, we plan to identify communities and compare them with those based on

This material is in part based upon work supported by the Office of Naval Research under Contract No. N00014-12-C-0043. The authors thank the Pacific Northwest National Laboratory's National Security Directorate for funding in part through the Laboratory Directed Research and Development program. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle under Contract DE-AC05-76RL01830.

sociolect use, but that is not our focus in here. As discussed below, we initially used the standard approach to assessing association in networks, called assortativity [14] [15] [16] and equivalent to Pearson's *r*. However, the assortativity statistic was strongly affected by group size[1], and it proved inadequate to our needs. We therefore developed two other new measures, explained in some detail below.

Explicitly identifying sociolects is a challenging problem in computational linguistics [7] [8]. For this effort, we took a less formal approach based on expert elicitation of topic-specific words. "Sociolect" term lists were constructed by identifying analytic topic areas of interest and interviewing appropriate experts. The results consisted of short list of keywords (typically around 20 terms) relevant and specific to the topic area, as judged by the expert. We call each such collection of words a *term list.* These constructed term lists are summarized in Table 1. Four topic areas were considered, which we refer to with the tags, *Narco, Syria, Bahrain,* and *Shooting.* As a control, we also constructed a random term list, which we call *Random*. For *Random*, a selection of 20 words was retrieved from the COCA (Corpus of Contemporary American English) online interface [17]. The COCA random retrieval selects one of the 60,000 most common words in the corpus with uniform probability. For the *Narco* term-list, the (Spanish) terms were taken from a glossary of narco-related terms published by the El Paso Intelligence Center [18]; thus it was developed by domain experts. All other term lists were elicited by us from subject-matter experts.

II. Data and Methods

A. Data

We will provide full details of the elicitations in a companion publication.

Name	Description	Language	Terms	Tweets Retrieved	Distinct Posters (Nodes/Edges)
Random	Control - COCA	English	20	1,127,895,980	20,325,708/30,411,053
Narco	Narcotics Violence in Mexico	Spanish	63	831,799	714,214/665,231
Syria	Syrian civil war	Arabic	20	27,159,300	237,538/615,397
Bahrain	Bahrain Shiite-Sunni conflict	Arabic	17	9,310,475	149,488/333,962
Shooting	Firearms Enthusiasts	English	28	279,829,564	6,894,309/6,196,719

Table 1: Summary of Term-Lists and Harvests

Twitter data were obtained from Pacific Northwet National Labs' suite of SociAL Sensor Analytics (SALSA). SALSA provides immediate access to and tools to manipulate data from over 20 billion entries in blogs, micro-blogs, comments, and mainstream news articles spanning 13-June 2011 through 11-March 2013. The system stores and indexes 140 TB of social media data in a distributed database, 60% in English and the other 40% among at least 60 languages. For this effort, data were collected exclusively from micro-blog Twitter, in which authors compose short messages, known as *tweets*, limited to 140 characters in length. Tweets are characterized by content, data, and author, among other fields.

A separate experiment was conducted for each term list in Table 1. For each term list, all tweets containing one or more words from the term list were retrieved from the SALSA corpus, creating a *term data-set* of tweets associated with the term list. Each tweet has a number of attributes, including text, author and post-time. For each unique author in a term data-set, the text of all tweets by that author were accumulated into an *author-specific text corpus*. Within each of these author-specific corpora, the number of distinct terms from the term list were computed, resulting in the author's *term-count,* or *TC*. This procedure yields a single, well-defined *TC* for each author for each of the five term lists. We refer to a group of others with the same TC as a *class*. We would like to understand mixing [15] [16] between different classes - qualitatively, our hypothesis is that authors who use "more" terms are more likely to interact. We make this statement more precise below.

Networks were constructed for each term data set as follows: each author in the term data set is associated with a distinct node in the network, and is characterized by its TC. Directed links between author/nodes were added based on mention (@-sign) and retweet (RT[2]) tags: an author/node (author-1) that mentioned or retweeted another author/node (author-2) in the author-specific text corpus (the accumulated tweets) would lead to an edge from author-1 to author-2. Edges are unweighted. In the event that author-2 did not exist in the term data-set, either because they were not in the full SALSA corpus, or had not used any terms from the term list, a node was added to the network for that author with *TC*=0.

We note that this simplifies the network by ignoring frequency of links and their temporal distribution. For example, a single link from author-1 to author-2 is used whether there are many mentions or a single mention. Extending the analysis with link weights based on mention frequency might reveal different phenomena, and we hope to explore this as the research continues. In this first analysis we intentionally avoid such complexities, although we believe that considering them could be useful.

Generally, classes with lower *TC* have many more nodes than classes with higher *TC*. This is intuitively reasonable: since filtering for a given term will select a fraction of the nodes in the corpus, repeatedly filtering for multiple (independent) terms will lead to a (geometrical) diminution

[1] *Modularity*, a closely related and widely used concept for detecting and quantifying community structure [24], has been similarly criticized . [20] shows algebraically that modularity is biased toward finding communities of approximately the same size. In response to efforts to address this problem by introducing a tunable parameter for community size, a recent paper [21] showed that modularity maximization, even with such a parameter, will either merge small modules which should remain separate or split large ones with should remain intact for networks with a broad distribution of module sizes (as found in many real-world networks). The authors even "conjecture that the tendency to simultaneously merge and split clusters is an inevitable feature of methods based on global optimization [21, p. 7], suggesting potentially widespread methodological problems in the area.

[2] MT (modified tweet) tags were ignored.

of group size. Occurrence of sociolect terms is *not* independent, as we discuss below, although class size still diminishes rapidly with increasing *TC*.

B. *Methods: Average Degree and Assortativity*

We begin by discussing node degrees in the constructed networks. We examine the set of networks associated with a term data-set, restricted to have *TC* higher than a given threshhold, *TCt*, generating t networks for each. For each such truncated network, average degree [19] was computed by dividing the total number of edges, N_E by the total number of nodes, N_N, yielding average degree as a function of *TCt*. We then plotted average degree for each term set for every value of t; note that the values of t for each term data-sets differed slightly.

All of these average degree plots, computed for both sociolects and the random term data-set, exhibited the same generic behavior: initial increasing average degree with increasing *TCt* up to a maximum, with subsequent decay. Maxima in all curves occurred around *TCt* = 6. However, we found no consistent differences between average degree curves from the sociolect-based term data sets and the random word term data set. We do note that the two Arabic-based term data-sets (*Bahrain* and *Syria*) yielded more highly peaked curves than the others. See Figure 1.

The peaked structure of the average degree curves implies some dependence of probability of connection with minimum term count; if *TC* were randomly assigned to nodes in a random network with connection probability p [14], average degree of the truncated networks would simply decrease as p^{TCt} . This may indicate the existence of a core group at around *TC* = 6, which may reflect strong consistencies in patterns of language use, perhaps due to word frequency effects, which we did not control for.[3]

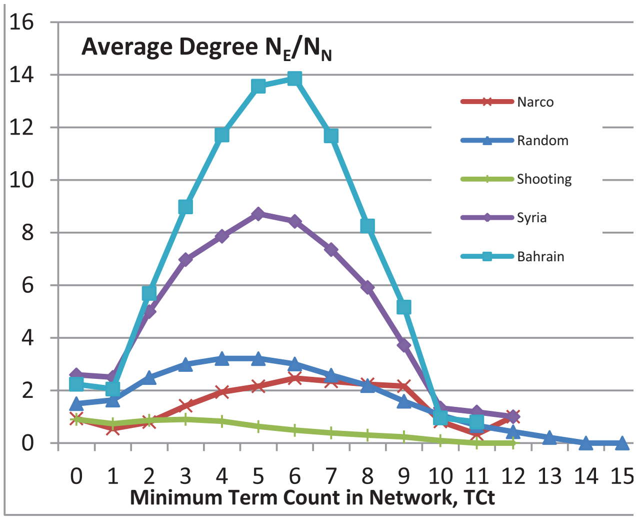

Figure 1: Average Degree vs minimum TC, TCt

A manifestation of homophily in social networks is *assortative mixing*, in which network nodes that share a similar property are more likely to be linked [11] [14] [19]. Newman [15] [16] provides a now-standard measure of assortative mixing[4]. Here, the assortative property of interest is the *TC*s of the nodes. To compute the metric one constructs a "mixing matrix" where each cell, N_{ij}, is simply the number of edges connecting nodes with *TC*=*i* to nodes with *TC*=*j*. This is a square matrix, symmetric for undirected graphs but not for directed graphs (such as the ones we used). The matrix can be normalized by total number of edges in the network, N_E, yielding the matrix $e_{ij} = N_{ij}/N_E$. (We note that this normalized mixing matrix includes no information about the number of nodes in each class, N_i.) This matrix is the probability that, given an edge between a source node *s,* and a destination node *d,* that the edge will connect a node with *TC i* to one with *TC j*. $e_{ij} = P(s \epsilon i, d \epsilon j \mid Edge(s,d))$, where *Edge(s,d)* is a Boolean function returning true if there is an edge in the network connecting node *s* to node *d*. As a probability, $\Sigma_{ij}\ e_{ij} = 1$, and one can speak of marginals over columns, $a_i = \Sigma_j\ e_{ij}$ and rows, $b_j = \Sigma_j\ e_{ij}$ which give the proportion of nodes from (for rows) or to (for columns) nodes with value *i* or *j*. The sum of row and column marginals equals 1 by construction. The *assortativity coefficient, r,* is:

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i} \qquad (1)$$

As Newman notes [15] [16], this is simply the Pearson correlation coefficient of the *TC*s paired by edges in the network. This makes intuitive sense, since the relationship of interest is precisely the association of values of the row and column values of a quantitative variable. We expected increasing association with increasing *TC*. However, the results of this initial assortativity analysis were not encouraging. Briefly, assortativity calculations yielded a coefficient close to zero, typically positive, but small. This held for networks constructed using tweets selected with the sociolect and random term lists. This also applied to subnetworks truncated by *TCt*. One explanation may be that Newman's *r* is very dependent on the size of the different classes in the network (strictly speaking, on the number of edges attached to a class, which will be the product of average degree and number of nodes). *TC* classes with large numbers of nodes will typically dominate the value of *r*. Due to the dramatic diminution of nodes (and edges) with increasing *TC*, the lower *TC* populations, which are less interesting from our perspective, dominate *r*.[5]

[3] Whether this reflects a core group of posters, a core group of terms, or a core size of term usage will be investigated in the next phase of the research.

[4] Although usually used to determine the relationship of node degree with link structure, assortativity can be used to characterize the relationship of *any* scalar node attribute with link structure, as Newman [15] makes clear.

[5] Since the focus of this paper is methodological, we note that we also performed an analysis where we computed the expected value of each cell in the normalized mixing matrix by simply multiplying the row and column marginals. We then divided the observed mixing matrix by the matrix of expected values, with the idea that the pattern of ratios should differ in the sociolect networks from

C. Methods: Edge Density and Term Independence

Motivated by this observation, we extended the analysis to include the node population information. We developed a simple formulation of the propensity of nodes in a network to connect: inverting the expression for $e_{ij} = P(s \epsilon i, d \epsilon j \mid Edge(s,d))$ gives:

$$P(Edge(s,d) \mid s \in i, d \in j) =$$

$$\frac{P(s \in i, d \in j \mid Edge(s,d))P(Edge(s,d))}{P(s \in i)P(d \in j)} \qquad (2)$$

where, $P(s \epsilon i) = N_i/N_N$, is the estimated probability of node s belonging to TC class i, and N_i is the number of nodes in class i. (2) introduces two important quantities that consider the *edge densities.* The first, $P(Edge(s,d) \mid s \epsilon i, d \epsilon j)$, is the probability of an edge, or interaction, between TC class i and TC class j. The second, $P(Edge(s,d))$, is the global probability of interaction in a network. (2) implies that $P(Edge(s,d) \mid s \epsilon i, d \epsilon j)=N_{ij}/N_iN_j$ and $P(Edge(s,d))=N_E/N_N^2$ [19]. This is intuitively reasonable: the probability of an edge occurring between two nodes in a network is given by the proportion of observed number of edges N_E to the *possible* number of edges in the network, N_N^2 (allowing for self-connection)[6]. Like average degree, this *edge density* is a global property of networks.

We plot, on a log scale, the edge density of subnetworks truncated by *TCt* in figure 2.

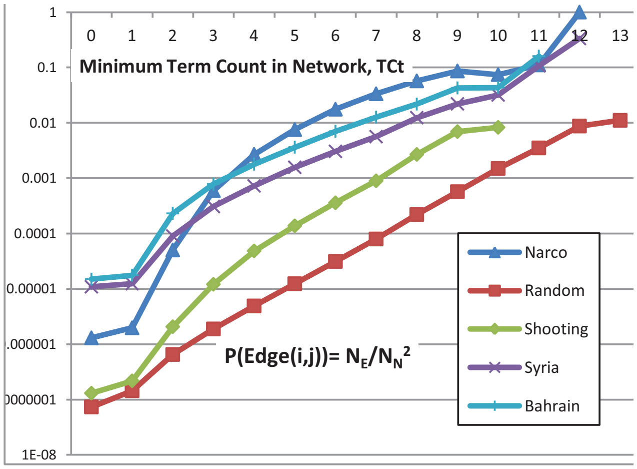

Figure 2: Edge Density vs TCt

This plot indicates that for a given number of terms, the probability of interaction in the sociolect networks is typically one to two orders of magnitude greater than the probability of interaction in the random network. This is most pronounced in the *Narco*, *Syrian* and *Bahrain* networks, less so for the *Shooting* network. This indicates that the shooting network is less connected – *i.e.*, has a lower density of edges – than the others. Whether this is a fundamental property of the underlying social system or an artifact of the informal sociolect construction process is a question for further study. We also note that all networks exhibit increasing interaction probability with increasing term count. Although the average number of links begins decreasing at *TC*=6, the number of nodes decreases more quickly; this shows the value of conditionalizing link probability on number of nodes, as done in equation (2). This may also be due to retweet effects, where, given several common words, the likelihood of nodes interacting through retweets increases, leading to increased edge density. Behavior of networks ignoring retweets is a topic for future study. Note that we do *not* regard this as an artifact of our methodology, since retweets are a fundamental information diffusion mechanism on Twitter. We have not controlled for word frequency, although the similarity of the slopes in the figure suggests that any such effects are similar across term lists.

Thus, we see a strong sociolect effect using this measure, which explicitly includes number of nodes in computing the conditional probability of links. This is similar to other work in which edge density has been used to identify subgroups [12], but we believe its use to identify assortative mixing is unique.

We have, thus far, presented two analyses: one, assortativity, based only on the edge structure of the network, N_{ij} (or equivalently e_{ij}), the other, edge densities, that adds in the node populations over TC classes, N_i. We now perform an analysis that depends *only* on the node populations N_i. Central to the definition of a sociolect is the notion that its constituent terms are not statistically independent. For example, someone who uses a domain-specific term, such as *breech, firing-pin* or *handload* for the case of shooting, should be more likely to use other domain-specific terms. The converse of this statement is that unrelated words should occur independently: given an arbitrary group of words, using one word from the group should not affect the likelihood of using another word from the group. Pursuing this reasoning, we define a quantity $Nt_i = \Sigma_{j \geq i} N_i$, the number of nodes in a network which have $TC \geq i$. We investigate the ratios $t_i = Nt_i/Nt_{i+1}$ - this is the fractional reduction in the node population when increasing the minimum number of keywords, *TCt*. If the probability of using a term, q_t, is independent of the number of terms used, this quantity should scale as q^{TCt}.

In Figure 3, we plot the ratio t_i/t_1^{TCt} . We note a striking difference between the ratios for the sociolect networks, which tend to increase with increasing term usage, compared to the ratio for the random network which remains O(1) until the last few terms, which have very small N's and are therefore noisy.

the control network. Although there was a tendency in some of the sociolect networks for higher values near the diagonal than for the network obtained using random terms, suggesting association, no reliable pattern was found.

[6] If self-connection is disallowed, then the possible number of edges between N nodes is $N(N-1)$ and the edge density of the network is $N_E/N_N(N_N-1)$. Our analysis is the same in either case.

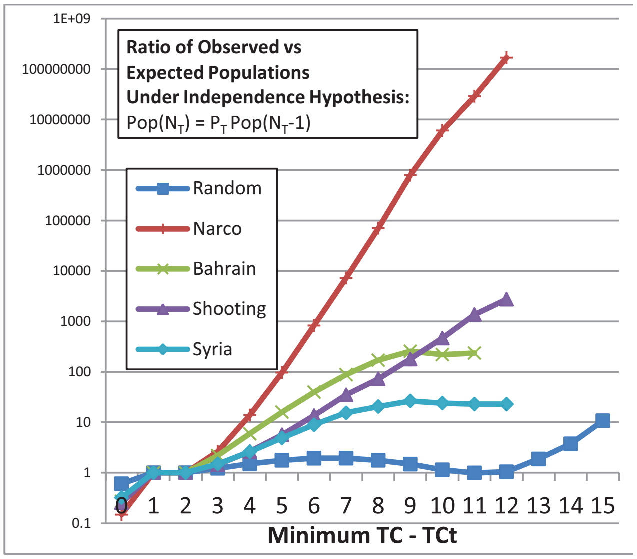

Figure 3: Term Dependence vs TCt.

III. Conclusion and Future Work

We applied a number of measures to look for signatures of sociolect usage related to linkage patterns in networks constructed by filtering billions of tweets based on various term lists. We believe that this focus is novel, and are not aware of any prior work specifically investigating the coupling between language and community structure [12] [13] [14]. Although we found increased levels of association between individuals who use common language, whether the terms used are from a sociolect or a random word list, the standard metric of average degree showed no reliable differences between sociolects and the control. Neither did Newman's assortativity [15] [16] nor a measure we developed using expected values for cells in Newman's mixing matrix differentiate between them. We speculate that this may be due to failure to account for large variations in sub-population sizes.

We observed that the node population as a function of term count dropped off dramatically with increasing term count. Therefore, we developed two novel metrics, one using both number of edges and number of nodes to compute edge density, and the other using only number of nodes to compute term independence. Edge density indicates that association for sociolect-based networks are typically two orders of magnitude stronger than association for random term list networks. Term independence shows a strong increase in likelihood for multiple term usage in sociolect-based networks compared to strong independence for term usage from the random term list.

In combination, these two methods provide a strong indicator of sociolect usage and associated community structure. We believe that considering node count, which crucially constrains number of possible edges, can substantially sharpen the analytical power of measures of association in networks more broadly. It explicitly addresses the issue of group sizes, which has been found to be important in the closely related problem of community detection [20] [21]. We are therefore quite encouraged that it will be possible to explore sociolect (and other social-media-based) networks more systematically going forward, using both these novel metrics and a range of others.

We plan to explore several additional methodological and substantive questions. Is it possible to find a parsimonious set of terms in a sociolect that can effectively differentiate a sociolect-using subnetwork from the larger network? Is it possible to find a sociolect without starting with an initial word set? Are retweets different from mentions in terms of sociolect usage and/or network structure? Can these methods be applied effectively to other social media, such as weblogs or Facebook? Are there ways to relate these networks to demographic variables? Can we use structural methods to find cores or key transmitters (or amplifiers) in networks, and do they have significant roles in information diffusion? We also intend to investigate the extent to which language and structure detects communities that are different from communities detected using traditional approaches based on structure alone; therefore, comparison of communities identified by multiple algorithms and the ones presented here should be enlightening.

We believe that refinement and application of these methods may be of considerable utility in identifying important subgroups in vast streams of social media data, with potential applications to military and related intelligence activities, marketing, political campaigns, dissemination of news and information, and the like. Multimethod approaches, such as the one presented here, are consistent with recommended best practices in social science research [22] and experimental design [23], and we are continuing to try to develop new methodological approaches to such problems. We also conjecture that community structure may best be characterized by a *vector* of measures, revealing multiple aspects of that structure, potentially relevant for different analytical goals.

Acknowledgements

We gratefully acknowledge Ivy Estabrooke of ONR for initial support and guidance of this work and CAPT Dylan Schmorrow of OSD and LT David Combs of NRL for ongoing guidance and support. We further acknowledge useful conversations with John Boiney, Barry Costa and Lisa Ferro of the MITRE Corporation. Finally, we acknolwedge the valuable contributions of Claudia (Eury) Gallegos.

References

[1] A. Rapoport, "Contribution to the Theory of Random and Biased Nets," *Bulletin of Mathematical Biology,* pp. 257-77, 1957.

[2] A. Rapoport, "Mathematical models of social interaction," in *Handbook of Mathematical Psychology Vol II*, vol. II, New York, NY, John Wiley and Sons, 1963, pp. 493-579.

[3] P. Erdős and A. Rényi, "On the evolution of random graphs," *Publications of the Mathematical Institute of the Hungarian Academy of Sciences ,* vol. 5, pp. 17-6, 1960.

[4] S. H. Strogatz and D. J. Watts, "Collective dynamics of 'small-world' networks," *Nature,* vol. 393, pp. 440-442, 1998.

[5] A.-L. Barabási and R. Albert, "Emergence of scaling in random

networks," *Science,* vol. 286, p. 509, 1999.

[6] R. Jakobson, "Linguistics and Poetics," in *Style in Language*, T. Sebeok, Ed., Cambridge, MA: MIT Press, 1960, pp. 350-377.

[7] M. Lewandowski, "Sociolects and registers–a contrastive analysis of two kinds of linguistic variation," *Investigationes Linguisticae,* vol. 20, 2010.

[8] M. Louwerse, "Semantic variation in idiolect and sociolect: Corpus linguistic evidence from literary texts," *Computers and the Humanities ,* vol. 38, no. 2, pp. 207-221, 2004.

[9] C. D. Corley, R. M. Farber and W. N. Reynolds, "Thought leaders during crises in massive social networks," *Statistical Analysis and Data Mining,* vol. 5, no. 3, pp. 205-217, 2012.

[10] "Twitter" http://www.twitter.com last accessed 23 April 2013.

[11] M. McPherson, L. Smith-Lovin and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology,* vol. 27, pp. 415-444, 2001.

[12] S. Fortunato, "Community detection in graphs," *Physics Reports,* vol. 486, no. 3-5, pp. 75-174, 2010.

[13] M. A. Porter, J.-P. Onnela and P. J. Mucha, "Communities in Networks" (February 22, 2009). Notices of the American Mathematical Society, Vol. 56, No. 9, 2009.

[14] M. E. J. Newman, "The Structure and function of complex networks," *SIAM Review,* vol. 45, pp. 167-256, 2003.

[15] M. E. J. Newman, "Assortative Mixing in Networks," *Physical Review Letters,* vol. 89, p. 2008701, 2002.

[16] M. E. J. Newman, "Mixing patterns and community structure in networks," in *Mechanics of Complex Networks*, R. Pastor-Satorras, J. Rubi and A. Diaz-Guilera, Eds., Berlin, Springer, 2003.

[17] M. Davies, "The Corpus of Contemporary American English: 450 million words, 1990-present," [Online]. Available: http://corpus.byu.edu/coca/. [Accessed 8 March 2013].

[18] El Paso Intelligence Center, "Language of the Cartels: Narco Terminology, Identifiers, And Clothing Style," U.S. Department of Justice National Drug Intelligence Center, 2010.

[19] S. Wasserman and K. Faust, Social Network Analysis: Methods and Applications, Cambridge: Cambridge University Press, 1994.

[20] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *Proc. Natl. Acad. Sci. USA ,* vol. 104, no. 1, pp. 36-41, 2007.

[21] A. Lancichinett and S. Fortunato, "Limits of modularity maximization in community detection," *Physical Review E ,* vol. 2011, p. 066122, 2011.

[22] J. Brewer and A. Hunter, Foundations of Multimethod Research: Synthesizing Styles, Thousand Oaks, CA: Sage Publications, 2006.

[23] D. Campbell and J. Stanley, Experimental and Qausi-Experimental Design for Research, Wadsworth Publishing, 1963.

[24] M. E. J. Newman, "Modularity and community structure in networks," vol. 103, pp. 8577-8582, 2006.