Original Research Article

BIG DATA & SOCIETY

Big Data & Society
January–June 2019: 1–15
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2053951719846634
journals.sagepub.com/home/bds
SAGE

Linguistically guided community discovery

Jean M Gawron[1], Alex Dodge[2], Ming-Hsiang Tsou[1], Brian Spitzberg[1] and Li An[1]

Abstract

Within some online communities, discussion often centers on issues on which writers take sides, and within some subset of those debate-prone communities, we find over time that particular sets of writers almost always end up on the same side of an issue. These sets we call factions. In this paper, we describe a tool to perform what we call faction discovery on online communities. Generalizing methods developed in the bibliometrics and information retrieval literature, we define a network determined by similarities of content in a community of users and add in direct evidence of online ties between users (e.g., link information such as mention-links). We then perform community detection on the network to find factions. Using a set of data collected from science and fantasy blogs, we show that the discovered factions accurately reflect an active conflict in the community leading to significant, politically related social fracture.

Keywords

Community discovery, Gamergate, linguistic features, Puppygate, social networks, unsupervised clustering

This article is a part of special theme on Social Media & Society. To see a full list of all articles in this special theme, please click here: https://journals.sagepub.com/page/bds/collections/social-media-society.

Introduction

Social media and social networks go hand in hand. Social media like Facebook, Twitter, Instagram, and Reddit work by allowing unfiltered user-generated content for users who know where that content will go. There are channels the content will follow; in many cases the exact set of users who will see the content is known; in others it is seriously constrained. The fact that the content will flow along known paths is part of the appeal, and the content itself is shaped by that fact. Content producers have an interest in producing content that will provoke a response among those who will see it. Users have an interest in joining networks that welcome what they have to say. Network and content are mutually reinforcing. In extreme cases, such as the human- and machine-generated messages exchanged at high speed in financial markets, knowledge of the network is indispensable to understanding the content (Christiaens, 2016). This mutually defining relationship between content and network is what motivates social media analysts to use both semantically driven and user-driven collection methods (Brooker et al., 2016). Semantically driven data collection includes any method that samples by topic, including sampling by hash tag or mention of a celebrity name. For instance, Tumasjan et al. (2010) collect political Tweets by tracking party names or selected politicians. The natural accompaniment for semantically driven collection is semantic analysis, including topic analysis (e.g., Cui et al., 2011) and document clustering (e.g., Wang and Kitsuregawa, 2004). User-driven data collection includes any user-based collection method. The simplest case would be collection for a set of users, but more often, user-driven collection proceeds by following the links in a network of users. For example, Sasahara et al. (2013) follows the retweet networks from a small set of users with a high volume of

[1] San Diego State University, USA
[2] NTENT, USA

Corresponding author:
Jean M Gawron, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA.
Email: gawron@mail.sdsu.edu

followers. The natural accompaniment of user/network-driven collection is link-based or structurally based analysis (Aggarwal, 2011), which includes algorithms for community discovery, node centrality, and influence analysis.

In this paper, we describe a tool to perform what we call faction discovery on online communities. Like social influence, factions are a hypothesis about the structure of a social group. One goal in hypothesizing factions is to explore the nature of ties between online users, illuminating differences between online ties and what social science researchers have traditionally called social ties (boyd and Crawford, 2012). An important feature of factions is that they persist over time; new users will naturally be assimilated into existing factions, and the identity of their faction can easily be determined by rerunning the relevant algorithms. Factions, then, are potentially dynamic modular units. As such, they may enrich the inventory of ways in which explanatory hypotheses about online communities can be ‘‘decomposed’’ using structural features of the data (Raghavan, 2014).

The method we propose for faction discovery is neither purely network-driven nor purely semantically driven, but rather a hybrid of the two. In contrast to some approaches which have applied network analysis to content analysis (e.g., Segev et al., 2015), this hybrid approach takes the nodes to be users (rather than memes or messages). In taking the network nodes to be users and in combining semantically driven analysis with network analysis, faction discovery resembles social influence analysis (Tang et al., 2009). It also resembles the kind of microgroup discovery described in Papakyriakopoulos et al. (2018), although the methods and goals of both differ somewhat from those described here. We begin by defining a network determined by similarities of content in a community of users we call a topic community. We then perform community discovery on this network to select sets of users we call factions, which are users united by the strong similarities of their content. Ultimately, we find that similarity of content alone is too weak a criterion for defining useful factions, and that similarity of content must be combined with direct evidence of an online tie between users (e.g., link information such as mention-links).

We first identify the kind of online community we will collect data from. We use the term topic community to refer to a community of users who produce texts more or less on the same topic and who read each other’s texts. A topic community may be a group of bloggers focusing on a particular political or social issue such as vaccine policy or climate change or they may be working in a particular academic area, or they may be contributors to online discussion fora on media of certain types, for example games or science fiction. The members of such topic communities not only read each other’s texts; they often write texts responding directly to contributions by another member. Texts and authors therefore are more or less important within the community according to how much attention they are paid. Within some topic communities, discussion often centers on issues on which writers take sides, and within some subset of those issue-prone topic communities, we find over time that particular sets of writers almost always end up on the same side of an issue. These sets we will call **faction**s. Given an online topic community, our chosen task is to find the factions within it, if they exist. Note that we are defining the notion faction with respect to conflict played out in online texts, but the concept of a faction as a group united by common purpose within a larger group has played an important role within social science in general (Dewan and Squintani, 2016; Persico et al., 2011; Zachary, 1977).

Suppose we treat all the productions of an author in the community as a single document. Then we can broadly characterize this problem as a document-clustering task. The most important novel feature of our approach will be to apply linguistically based document clustering methods to the problem of faction discovery. We note that, at least for many of the online topic communities we see, satisfactory solutions of the faction problem should be able to show how the factions can be grouped so as to divide the community in two, because most issues have two sides, and we find with great consistency the same sets of people showing up together on one side of an issue. Many clustering algorithms are hierarchical and will do this for us automatically (Manning et al., 2008). That is, they either work by successively merging small clusters until a two-way partition is found (agglomerative) or by successive splits of an initial two-way partition (divisive).

Many document-clustering methods are similarity based. That is, they define some kind of document similarity relation, and they employ a clustering algorithm to assign the documents to clusters based on how similar they are. We can distinguish similarity-based approaches to document-clustering along two dimensions: (a) how similarity is defined and (b) what algorithm is used for clustering.

With respect to (a), we can distinguish link-based versus term-based (or linguistically based) similarity approaches to document clustering.

Link-based approaches start with some kind of link between texts that can be identified by a fairly superficial (crucially nonlinguistic) analysis, for example, hyperlinks between web pages or bibliographic citation. Once a linking relation is chosen, two basically independent kinds of similarity between documents are definable, similarity based on in-links (co-citation) or similarity based on out-links (co-reference or bibliographic coupling). Two documents have high in-link similarity if they are pointed to by the same documents. Two documents have high out-link similarity if they point to all the same documents. These are two different notions of similarity which a large body of work has shown to be only weakly correlated (Kleinberg, 1999; Small, 1973).

Term-based analysis requires some kind of linguistic analysis. Words or phrases or sentences are extracted, and a linguistic representation is constructed. Document similarity is then defined in terms of how similar the linguistic representations are. Within term-based approaches, we can also distinguish two approaches, differing in where the language comes from: document-content based and citation-context based approaches. One can define the documents most similar to a document D based on the language IN D (document-content based) or based on the language from other documents used to refer TO D (citation-context based). Aljaber et al. (2009) explains citation contexts as follows: "Citation contexts refer to textual descriptions of a given scientific article found in other articles in the document collection which cite it."

In this paper, we will develop a novel term-based technique for document clustering which has some of the features of a link-based approach. Because this is basically a feasibility study, we will experiment only with analyzing the document content, though nothing in principle prevents applying the same idea to a citation-context approach. The core idea, motivated by the need to focus on information most likely to be relevant to faction discovery, is to limit our analysis to terms used to refer to people, organizations, and other works, including other documents or authors in the topic community. We will provide some theoretical background for this choice in the "Approach" section. For the moment, note that once we limit our attention in this way, the distinction between the two kinds of term-based approach quite resembles the distinction between the two link-based approaches. Our linking relation is linguistic reference. Content-based similarity is like similarity based on out-links; we are defining the similarity of documents A and B based on the similarities of the sets of entities referred to (linked to) in A and B. Citation-context based clustering is like clustering based on in-links; we are defining the similarity of documents A and B based on the similarities of the sets of entities that refer to (link to) A and B.

The other dimension along which similarity-based approaches differ is in the choice of a clustering algorithm. To cluster documents, we will employ a class of algorithms known as community discovery algorithms, which will require building a weighted graph representing relationships among documents. There are a variety of advantages to such an approach that we will try to describe in the "Approach" section, but the most essential feature is that there is a class of community discovery algorithms which are hierarchical, so they will serve to describe the way in which the community is polarized into two sides. To our knowledge this is the first example of using this class of algorithms in the context of term-based document clustering, though it is not the first time they have been used for faction discovery.

Because our approach to defining term-based similarity is so similar in spirit to a link-based approach, and because our data consists of blog posts with hyperlinks, it is natural to explore combining a link-based approach and a term-based approach, particularly since such hybrid systems have been shown to be more effective in applications such as document retrieval. We will be using many of the insights from previous work in determining how to effectively integrate the two kinds of information into a single similarity measure; to our knowledge, this is the first time that combining link-based and term-based similarity has been used in the context of a graphical clustering algorithm. That creates some technical challenges because the algorithms we use require undirected graphs; we will sketch a solution in our "Approach" section.

Summarizing the basic features of our term-based approach: we represent a document using only referring language, specifically proper names referring to people, organizations, and documents; we leave the more general task of dealing with referring noun phrases (NPs) such as "my worthy opponents" to later work. We define document similarity for such representations and use it to define a graph, and we apply community discovery algorithms to find the factions. We will also experiment with combining link-based and term-based information.

The rest of this paper proceeds as follows: We discuss the relationship of this application of document clustering to previous work. Then we describe our approach, identify a dataset, and describe a particular experiment to test it, presenting results that show the linguistic information helps. Finally, we discuss the analytical consequences and sketch some directions for future work.

Previous work

The root of the tree from which the current study springs is early work in bibliometrics on citation graphs, graphs representing the citation relation between academic papers. Citation graphs have been an important tool in the study of research communities and academic papers ever since large accurate citation databases like the Science Citation Index came into being in the 1960s (De Solla Price, 1965; Garfield, 1955).

The usefulness of citation graphs in understanding the connectedness of scientific ideas was first pointed out in Small (1973). From our present perspective the most important aspect of Small's paper is its introduction of similarity relations. Beginning with the basic linking relation x cites y, Small derives two similarity relations, co-citation (similarity of citer sets) and co-reference (similarity of papers cited). Choosing to focus on co-citation, Small builds a weighted network in which the weight of the link between documents i and j corresponds to their co-citation strength, (that is to the degree of similarity in the sets of papers citing them). We will refer to any graph in which a link weight represents some measure of the similarity of two documents as a similarity graph.

Since Small (1973) there has been a large body of work exploring the clustering and classification of documents using co-citation, for example White and McCain (1998), Aljaber et al. (2010), Zhao and Strotmann (2014). See Chen et al. (2010) for an excellent survey. In many cases an approach combining co-citation and co-reference has been shown to be useful (Kleinberg, 1999; Zhao and Strotmann, 2014).

We turn to approaches that use linguistic information, which we called term-based approaches in the ''Introduction'' section. Term-based representations of document content have been standard since the earliest information retrieval (IR) systems (Salton, 1971; Salton and McGill, 1983) and are still in use in modern IR algorithm such as Google's. Thus, it was quite natural to investigate document clustering of document-content representations in various related applications, for example unsupervised discovery of topics in a document set (Wang and Kitsuregawa, 2004), preclustering a corpus and retrieving clusters rather than documents (Salton, 1971), and cluster-based navigation (Cutting et al., 1992; Weiss et al., 1996). Using a citation-context approach, Bradshaw (2003) retrieves documents by indexing citation-context.

Many researchers have explored combining information from different types of document representation. In web retrieval studies, Haveliwala et al. (2002), Zhao and Strotmann (2014), and Ritchie et al. (2008) achieve improved performance by combining citation-context representations with document-content representations. Robertson et al. (2004) discusses the general problem of combining multiple representations of a document to improve the performance of IR systems, for example using information from various text fields, such as title, abstract, and text body. Aljaber et al. (2009), in a text-clustering task exploring various combinations of link-based and term-based approaches, achieve their best results with a term-based approach that linearly combines similarity scores for citation-context representations and document-content representations, along the lines of Robertson et al. (2004).

The idea of a hybrid system combining link-based information with term-based information, as we do here, has also received attention. For a web retrieval application, Weiss et al. (1996) cluster document using a similarity measures that combine content-based and link-based similarity, by taking the maximum of two similarity scores. Wang and Kitsuregawa (2004) linearly combines text similarity scores with link similarity scores. We pursue a similar strategy below.

Approach

Our task is faction discovery using document clustering. Two choices described in the ''Introduction'' section deserve brief discussion. First, why restrict document representations to referring terms that denote people, organizations, and cultural products? Second, why use a graphical representation and graphical clustering algorithm?

We first discuss the motivation for reference-based document representations. Since our problem is finding factions, and since linguistic features in general are noisy and diffuse, our goal is to zero in on a set of features that will be of particular help in identifying factions. Seminal results in social psychology tell us that factions evolve as an US–THEM mentality evolves (Anderson, 2003; Gupta, 2008), characterizing one set of people simultaneously as the outsiders (not US) and as a threat (THEM). In terms of online discussion of issues this means characterizing people in terms of the groups they belong to or are perceived to belong to instead of their positions (liberals, conservatives). Studies of online hate groups have shown that patterns of US–THEM reference can reliably capture the degree of militancy of online hate groups (Gawron et al., 2012). We thus hypothesize that the sets of referents and the linguistic choices of how to refer will provide many reliable indications of faction membership. One obvious example is the characteristic use by white hate groups of derogatory and offensive terms for nonwhite groups, but less extreme examples can be easily found. Adamic and Glance (2005) report a number of interesting cases in which references to particular political or media figures are much more likely to be made by blogs of one political orientation than the other. For example, right wing blogs were much more likely to refer to Dan Rather, Yasser Arafat, Michael Moore, and Howard Dean than left wing blogs. Left wing blogs were much more likely to refer to Dick Cheney, Colin Powell, Karl Rove, and Tim Russert than right wing blogs. Note that many of these characteristic references are out-of-faction reference (references to THEM).

The second question is why resort to a graph-based approach to faction discovery? First, we believe that topic communities will have structure typical of social networks. Such networks are not random. Their edge distributions reveal high levels of organization: ''with high concentrations of edges within special groups of vertices, and low concentration between these groups" (Fortunato, 2010). This clustering property is called community structure (Girvan and Newman, 2002). We hypothesize that in an appropriately defined graph of online topic community, factions will reveal themselves as smaller communities.

We turn to the problem of choosing a community discovery algorithm from among the large field of competitors. We began by eliminating algorithms that discovered communities in dynamic networks, because the size and timespan of our datasets would not allow us to appropriately treat the evolution of the networks in time. We also eliminated approaches that allowed overlapping factions, because socially, an individual is required to choose among factions. In preliminary pilot studies, we had greater success with modularity-minimizing approaches such as the Louvain algorithm in Blondel et al. (2008) and the fast greedy algorithm of Clauset et al. (2004) than with approaches based on maximizing flow within communities using the map equation as is done in Rosvall et al. (2009) or approaches based on label propagation (Raghavan et al., 2007), or approaches based on Random Walks (Pons and Latapy, 2005). Considerable improvements for label propagation are reported with a label-boundary approach as in Gui et al. (2018), but we have not yet experimented with this version.

There is considerable intuitive appeal to idea of the best communities being those that minimize modularity. The task of a community discovery algorithm is to automatically identify such community structure by partitioning the network in ways that minimizes the weight of inter-partition links, while maximizing the weight of intra-partition links; this language roughly describes the property called modularity (Blondel et al., 2008; Newman, 2006a). Blondel et al. and Newman both present hierarchical community discovery algorithms that seek to optimize modularity; we will refer to the Blondel et al. (2008) algorithm as Louvain (as it is commonly known) and to the Newman (2006b) algorithm as Newman. Given the graph in Figure 1(a), either of the two algorithms will partition the graph as in Figure 1(b), because there are only two inter-community edges, and no other way of partitioning the graph has fewer inter-community edges.

It is a remarkable fact that such algorithms, which rely exclusively on the link structure of the network, can be quite revealing when applied to social networks. Community detection can discover the highly interdependent parts of the network, such as the parts of an organization devoted to particular functions, or infer hierarchical structure (Clauset et al., 2008; Qiu and Lin, 2014), or find weaknesses that can predict possible fracture. Figure 2 shows the results of applying community discovery to data from a famous study of the emergence of factions, Zachary's (1977) karate club. The weighted links represent degree of joint social activity outside the club, as determined by one-on-one interviews. The factions determined by Zachary's fieldwork are shown in panel (a) of Figure 2.

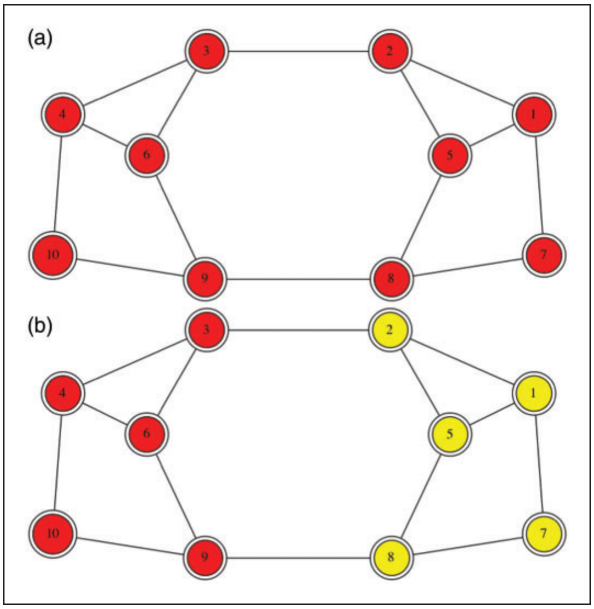

Figure 1. Community discovery algorithm operating on a simple graph.

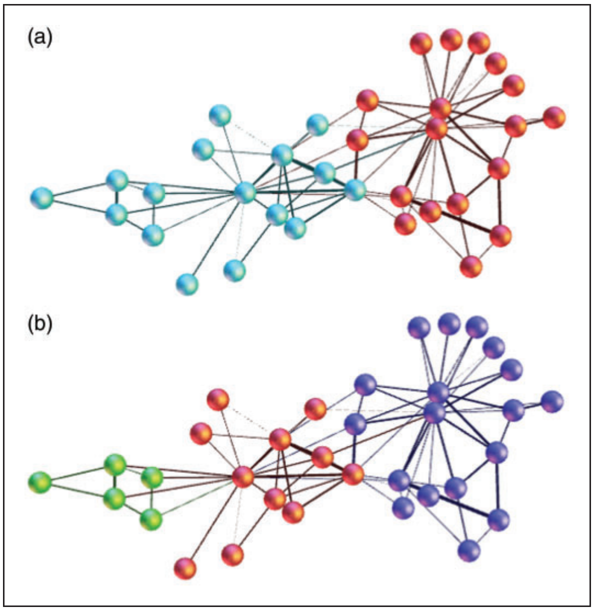

Figure 2. Panel (a) shows the factions proposed by Zachary in his famous karate club study, while panel (b) shows the result of applying Newman's community detection algorithm to the graph.

The result of applying the algorithm is in panel (b). Note that the algorithm discovers three communities, not two, but that the communities are 100% pure, that is none of the proposed communities mixes members from different factions. Thus, one of Zachary's factions has been reconstructed exactly, while the other has been split in two. However, since the algorithm is divisive, we can recover the first split, and that is indeed a split between Zachary's two factions.

Beyond what the communities themselves reveal about the function of the entire network, there is also fact that each faction discovered will be a subgraph with its own structural properties. For example, we can apply various measures of centrality to ask who are the most central actors in the faction. This is not possible with a nongraphical approach to document clustering.

Thus, an important motivation for using a graphical approach to factions is that we would like to study them as social networks. We believe the global goal of the algorithms described in the "Introduction" section (maximizing modularity) makes social sense, and if our similarity relations are valid indicators of social closeness, we can study the structural properties of the network we build and gain social insights.

We turn then to the question of how to represent our problem as a graph. Our starting premise is that we have collected data about a group whose members produce documents, together with overt links between the documents which naturally form a graph. In our case, the links between documents are hyperlinks. We describe an approach that measures the similarity of documents by combining link-based and term-based information and uses that similarity information to define a graph. We begin with the simpler case of link-based information.

As pointed out in discussing Small's work, Small did not represent links directly in a graph; instead he represented link-based similarity in a similarity graph. Applying that idea to hyperlinks, we construct two distinct similarity graphs, one for in-link similarity and one for out-link similarity. In the graph representing in-link similarity, which we call CoCit, a link connecting document i and document j in the graph has weight μ (a number between 0 and 1) if the sets of documents linking to document i and document j are similar to degree μ (among the various ways of calculating the similarity of two sets, we use cosine below). In the graph representing out-link similarity, which we call CoRef, the weight of the link between i and j is μ just in case the sets of documents linked to by documents i and j are similar to degree μ. We compute the similarity of CoRef and two CoCit sets the same way, using cosine. Note the difference between either of these two similarity graphs and a graph which directly represents the linking relations (which we call a links graph). In the links graph, a link from i to j is directed and unweighted and means i cited j. In the similarity graph, a link between i and is undirected and has a weight μ and means that i and j are similar to degree μ. Note, for example, that i and j can have strong similarity in the CoRef graph without either one citing the other, if they cite many of the same papers.

Turning to term-based information, we adopt the classic bag-of-words representation of a set of documents (Manning et al., 2008; Salton and McGill, 1983) in which N documents using a vocabulary of V words becomes an N×V term document matrix. As argued above, the vocabulary features of a document should include only proper names describing persons, works, and organizations. This is not exactly a bag of words because names may have arbitrary numbers of "words" in them, so the features in this features set may be unigrams, bigrams, or even 4-grams, as, for example in "Secretary of State Clinton." Some care is taken to do "name-stemming" and merge name variations referring to the same person (e.g., "Heinlein" would be merged with "Robert A. Heinlein"), but references by first name alone are not merged with longer references. Thus, "Hilary" would not be merged with "Secretary of State Clinton." We think this is correct because those ways of referring have very different connotations. Features are weighted using pointwise mutual information (PMI) (Church and Hanks, 1990). PMI is a measure of how unexpected the frequency of a feature in a document is, given its overall frequency, and thus how informative the feature is. Documents are represented using only the F most frequent names, where F is a parameter we set by training. We write the similarity score between documents i and j as LingCoRef(i, j), and compute it in the usual way, as the cosine of the two document vectors. In the LingCoRef graph, a link of weight w between two documents means they are similar to degree w in the sets of significant proper names they use.

We do not build a single representation for the linguistic information and the link-based information. Rather, following Robertson et al. (2004), we use CoCit, CoRef, and LingCoRef to compute three similarity scores. We then define the overall similarity of two documents as an unweighted linear combination of all the similarity components, as did Wang and Kitsuregawa. That is, we compute Sim(i, j), the overall similarity of two documents, by just adding the component similarities

$$\mathrm{Sim}(i,j) = \mathrm{CoCit}(i,j) + \mathrm{CoRef}(i,j) + \mathrm{LingCoRef}(i,j) \tag{1}$$

Once Sim is computed, the final step in defining a similarity graph is to choose the maximal number of neighbors a document can have. In the maximal similarity graph, there is a link between any two documents, with a weight equal to the similarity of those documents, including weights of zero. In our implementation, each document is linked to its K most similar neighbors, for a value of K to be set by training.

The use of similarity graphs helps in two ways. First it gives us a very natural way of combining term-based and link-based information. Without transforming the link-based information into similarity scores, there would be no obvious principled way to combine the linguistic information. In addition, using the similarity graph addresses a significant technical challenge in running community discovery. A general limitation for graph-based clustering is that it works only on undirected graphs. Thus, in order to use community discovery algorithms like Louvain and Newman on a graph directly representing citation links, one must ignore any directionality in the links, essentially erasing the distinction between A citing B and B citing A. The distinct similarity relations co-citation and co-reference naturally represent this difference, yet still produce undirected similarity graphs. Figure 3 illustrates the idea. Arguably, all the vertex pairs in Figure 3(a) have similarity 0. They share neither citers nor works cited; on the other hand, in Figure 3(b), A and C are similar in citing B, and B and C are similar in both being cited by A. A and B are not similar. Treating graphs (a) and (b) as undirected reduces them to the same graph, obliterating the differences.

Figure 4 gives the similarity graphs derivable from the links graph in Figure 3(b). There are three distinct graphs, a CoCit graph, a CoRef graph, and the summed similarity graph, CoCitRef. The third graph clearly represents the similarities noted above, while remaining undirected. A is connected to C, C is connected to B, and A and B are unconnected.

Summarizing the point of this example, constructing a similarity graph allows us to accurately represent information obtained from directional links while still performing community discovery on an undirected graph. There is, however, an empirical question that this move raises: Will community discovery work as well on the links graph as it does on the corresponding similarity graph? Might it even work better?

For purposes of illustration, and to introduce our evaluation method, we will apply the two community discovery algorithms to two similarity graphs constructed from nonlinguistic data. In these examples there will be no LingCoRef scores. Consider first Zachary's karate club, in which the links represent joint out-of-club activity, and hence are naturally undirected. We first transform Zachary's links graph (see Figure 2) into a similarity graph in which the similarity score of two actors is defined as the similarity of their neighbor sets (see Figure 5). We then run the two community detection algorithms on that. In fact, as shown by the colors in Figure 5, both Newman's algorithm and the Louvain algorithm propose two communities which exactly reconstruct Zachary's factions. This result is arguably better than the result shown in Figure 2(b), if the goal of faction discovery is to predict fracture as Zachary does; the scoring metrics discussed below support this evaluation.

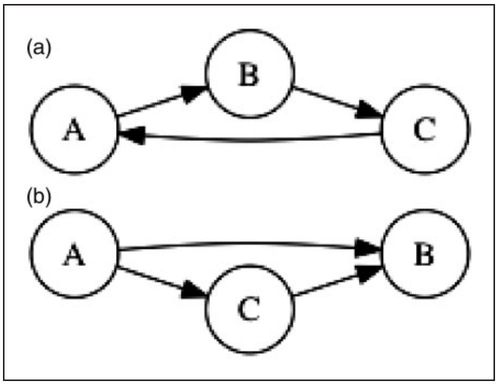

Figure 3. Ignoring the directionality of the links by treating all links as undirected can lose important similarity information. In graph (a), none of the nodes have any similarity properties; in graph (b), A and C co-refer; and B and C are co-cited.

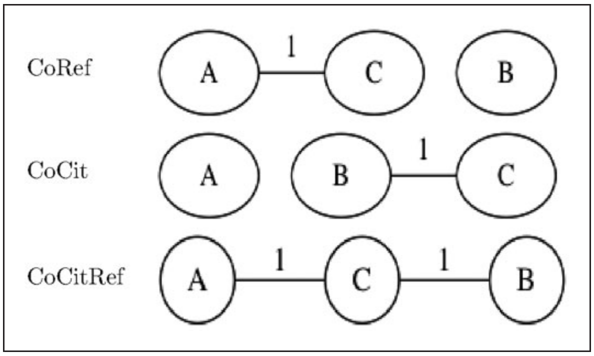

Figure 4. The similarity graphs derived from the links graph in Figure 3(b), using number of cociters or co-references as the similarity measure. The third graph is the sum of the first two.

Figure 6 displays the contrast between a links graph and a similarity graph on a much larger-scale example, the Polblogs data of Adamic and Glance (2005). In this case the links represent hyperlinks and are intrinsically directed. Thus, there is a trade-off in using the community discovery algorithms on the links graph. The directionality of the links must be ignored. On the other hand, information about directionality can be preserved in the similarity graphs, by separating in-link (CoCit) and out-link (CoRef) similarity. This dataset has no associated texts; thus, only CoCit and CoRef information has been used. The colors of the links

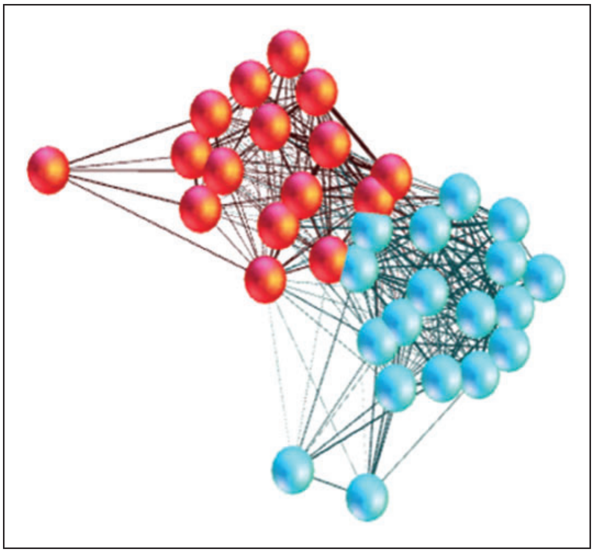

Figure 5. Zachary's karate club as a similarity graph. The colors represent communities assigned by Newman and Louvain and exactly match Zachary's factions.

graph (top) reflect the actual political orientations assigned to the sites pictured and show that the link structure is sufficiently informative for the layout algorithm to do a good job of separating orientations. The lower graph is the similarity graph summing co-citation and co-reference relations (the colors shown reflect community assignments by Louvain). Note that the similarity graph is much more densely connected, and although the layout algorithm still seems to be separating two large groups, the boundary between the two groups would be much more difficult to find without the aid of the colors.

Nevertheless, when we do the numbers, it turns out that the two graphs capture very nearly the same information. Table 1 presents the result of applying a similarity component approach to the Polblogs data. The table shows three types of similarity graph results, one based only on co-citation similarity, another based on co-reference only, and a third based on the sum of the first two kinds of similarity (the row labeled CoCitRef). Table 1 also gives the results of using the algorithms on the links graph, risking the kind of information loss illustrated in Figure 3.

The figures shown in Table 1 evaluate how well the community detection algorithms reconstruct, or at least are compatible with, the known political orientation of the blog sites. The first figure shown in each row is accuracy (or weighted purity), which counts a cluster assignment for individual x as correct if the majority of the membership of that cluster belongs to the same group as x. The trouble with accuracy is that it does not correct for clustering success due to chance, which will vary with the sizes of the clusters and their number.

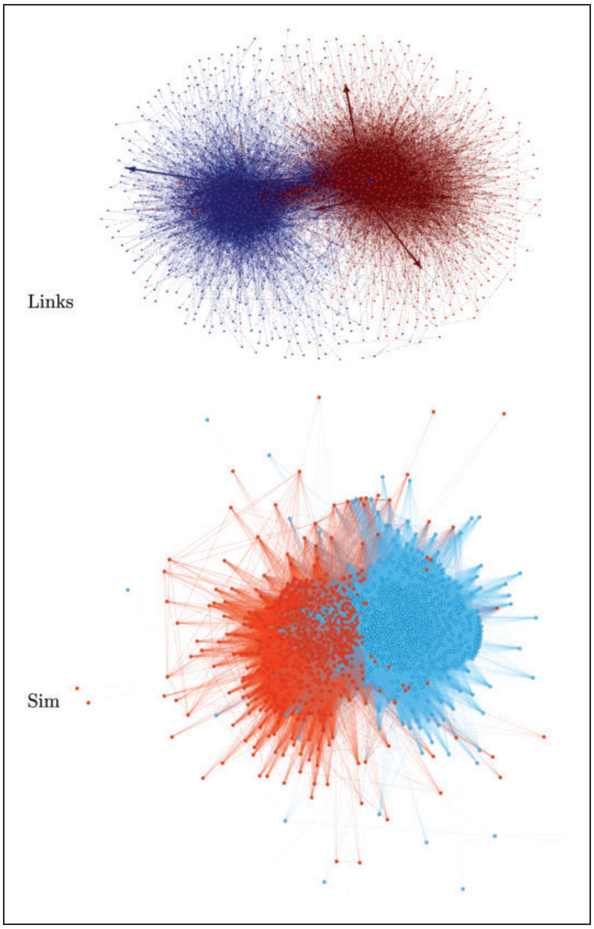

Figure 6. The Links graph (top) and the summed similarity graph (bottom) for the Polblogs data of Adamic and Glance. Note the marked increase in edge density in the similarity graph.

Table 1. Adjusted mutual information (AMI) and purity scores for Louvain and Newman community detection algorithms on the links graph and three similarity graphs for the Adamic and Glance (2005) Polblogs data.

		Purity	AMI	Communities
Louvain	Link	94.76	.559	9
	CoCit	87.70	.456	4
	CoRef	82.30	.254	5
	CoCitRef	94.76	.562	9
Newman	Link	95.25	.727	2
	CoCit	87.32	.511	2
	CoRef	82.40	.267	3
	CoCitRef	95.17	.724	2

At one extreme we have one super cluster, which for a politically balanced sample of N sites, will earn an accuracy score of 50%, and at the other, N clusters of size 1, which will earn an accuracy score of 100%.

Intuitively, both ways of clustering the graph are maximally uninformative and should earn scores of 0. Accordingly, we use Adjusted Mutual Information (Vinh et al., 2010), which is based on Mutual Information (Cover and Thomas, 2012), a standard information-theoretic measure of the relatedness of two distributions. AMI is Mutual Information adjusted for success attributable to chance. The scale is from 0 (a clustering whose success is due entirely to chance) to 1 (a clustering that exactly captures the known classes). The score assigned to both the 1-cluster and the N-cluster solutions above is 0. Zachary's karate graph provides a very nice illustration of the difference between AMI and Purity; the community assignments for the karate club graph in Figure 2(b) earn a purity of 100%, but an AMI score of only .693, because AMI penalizes a solution that splits one of Zachary's factions in two, while the two-community solution found by applying community detection to the similarity graph in Figure 5, which reconstructs both factions exactly, gets a purity score of 100% and an AMI score of 1.00.

According to the evaluation scores in Table 1, summing the two similarity graphs achieves better results than either similarity graph alone, and for both algorithms, the summed score is practically indistinguishable from the results obtained using the links graph (.724 versus .727 for Newman, and .562 versus .559 for Louvain). Clearly, in this dataset, the summed co-citation and co-reference similarity graphs contain very much the same information as the links graph. In general, however, the two methods are not equivalent. We will show that for the science fiction and fantasy blog data in the "Data" section, the summed similarity graph produces significantly better clustering results than the links graph, even without added linguistic information. We hypothesize that summing different similarity components helps faction discovery under two conditions: (a) the summed information is relevant to establishing social ties; and (b) the summed components are largely independent. In his original co-citation paper, Small argued that co-citation and co-reference similarity correlated poorly, but he based this on the statistics for research papers. In the Polblogs domain, 24.3% of all links are reciprocated, suggesting that you-link-to-me-and-I'll-link-to-you is a very common practice, if not exactly a norm. This proliferation of bidirectional edges in the network may well neutralize the difference between the summed similarity graph and the links graph.

Turning to the systems we will compare: We will compare three different systems, Link, Ling, and LingLink. All three systems will use community detection on similarity graphs. The Link system will use a graph with only link-based information, the sum of CoCit with CoRef. The Ling system will use only term-based information; the similarity graph will contain only similarity scores for the sets of proper name references in two documents. The LingLink system will combine the similarity scores from the Ling and Link systems.

Data

The data for our case study was collected in July of 2015 from 218 sites in the science fiction and fantasy blogging community with the aim of tracking an ongoing fracture in the community known as **Puppygate**. The history of the fracture will briefly be described but it will be useful to first give some background for a more famous fracture which both shared many features with Puppygate and strongly influenced it.

In August 2014, a number of posts attacking prominent women in the video gaming industry appeared in social media devoted to gaming. The women being targeted included cultural critic Anita Sarkeesian (creator of the website "Feminist Frequency" and frequent critic of the sexist slant of video games), Zoe Quinn (co-creator of Depression Quest), and Brianna Wu (game developer and journalist). The attacks were very broadly based, appearing on Internet Relay Chat channels, Twitter, Reddit, 4chan, and 8chan. They were inspired in part by the release of video games such as Depression Quest exploring darker real-life themes and challenging the traditional role of video gaming as pure escapism; participants often saw themselves as reacting against a strain of "political correctness" that had "infected" the gaming world, but the attacks also included a personal component. A former boyfriend of Zoe Quinn posted a highly critical piece alleging that Quinn had exchanged sexual favors for favorable coverage of her game. A component of the reaction that followed was a discussion of journalistic ethics in video game coverage, but a darker strain emerged as gamers posted threats of physical violence against Quinn, Sarkeesian, and Wu (Colbert, 2014; Parkin, 2014a, 2014b). In a later phase, the controversy took on an economic dimension as critics of Gamergate successfully pressured advertisers to pull their ads from game sites (Wingfield, 2014).

Thus, Gamergate is the story of political conflict being played out in a loosely connected online "world" not created for political purposes, but suddenly swept up in what conservative critic James Davison Hunter (1992) labeled "the culture wars." Gamergate has become a classic, rather well-publicized example of how polarization in the political sphere begins with and plays out in the cultural arena, attracting coverage from *The New Yorker, The New York Times*, and Stephen Colbert. Gamergate also illustrates how serious the consequences of such eruptions can be, inflicting serious economic harm, creating what appears to be a permanent community fracture, and inspiring death threats. Crucially, Gamergate was described by the players themselves as a reaction against perceived political correctness, against pious “Social Justice Warriors.”

All of these features arise in the Puppygate controversy. The gaming community is replaced by the science fiction and fantasy community. Defining issues specifically related to feminism are replaced by general “social justice” issues. But the theme of a reaction against perceived political correctness is still a strong component, and fractures created in the community appear to be equally serious.

The Puppygate controversy concerns the alleged hijacking of the science fiction community’s annual award (the “Hugo” Award) by a group of conservative writers, fans, and publishers who circulated slates of eligible works, urging their readers and followers to read and consider nominating those works. Two slates in particular, known as the “sad” and “rabid” puppies slates, attracted considerable attention in 2014, earning a number of nominations; and in 2015, after a similar campaign, 18 of the 20 award finalists came from one of the two “puppy” slates. Angry critics like John Scalzi (2015) advised voting “No Award” when all the nominees were substandard, joined by *Game of Thrones* author George R. R. Martin (2015), and both became key figures in the anti-puppies reaction. Indeed, No Award won in many categories. The controversy took on a “Gamer-Gatish” color because the “puppies” (in particular “sad puppy” author Larry Correia) claimed they were reacting against a cabal of liberal “Social Justice Warriors” in control of the Hugos, casting political correctness in the awards process as their main target. Correia (2015) wrote: “This is just one little battle in an ongoing culture war between artistic free expression and puritanical bullies who think they represent real fandom.”

The most obvious similarity between Gamergate and Puppygate is the anger on both sides. Many of the texts in the data sample are vitriolic; personal attacks are commonplace, and the fact that the controversy also attracted the attention of major media players such as *The Wall Street Journal* (Rapoport, 2015), *Entertainment Weekly* (Biedenharn, 2015), *The Guardian* (Flood, 2014), and *The Telegraph* (Nobyline, 2015) probably made things worse. Although the media coverage of Puppygate was less intense than that of Gamergate, as would be expected for a battle in which the size of the market at stake is roughly an order of magnitude smaller, it is clear that the controversy attracted attention in venues that do not make coverage of the science fiction marketplace a regular practice.

The Puppygate controversy provides an ideal test case for our approach. First, it satisfies our structural requirements: It contains online texts produced by a community in fracture, with hyperlinks that can be used to infer social ties; moreover, the blog documents themselves very clearly establish factional affiliation, and thus in principle contain information that should be of help in identifying factions. Finally, the science fiction community is not a topic community created to discuss a political or policy issue with two predetermined sides. Rather the factions that emerge from the Puppygate data are formed as writers who have generally been discussing nonpolitical topics take sides in a debate that has suddenly become politically charged.

An additional feature of interest is that cross-factions links are quite common, often because bloggers embroiled in the debate are quoting sentiments they abhor. Almost exactly 2/3 of the hyperlinks are between nodes on the same side of the controversy (758 out of 1142), but that means 1/3 span between communities, making it more difficult for a community detection algorithm operating on link information alone to identify the factions. Compare this with the two cases in which our community detection algorithms achieve a high degree of success, the Polblogs data, in which 91% of the links fall within community, and the accuracy of the best algorithms was over 90%, or the karate club graph, in which 87.2% of the links are within faction links.

Finally, the Gamergate and Puppygate fractures are excellent examples of the practical utility of faction discovery. Similar sorts of right/left fracture can be found in many online communities. There is a similarity in issues and similarity in arguments in many confrontations between two factions, reflecting a growing political polarization in online communities. More specifically, the continued relevance of the “social justice in media” debate exhibited in both Gamergate and Puppygate can be seen in the more recent trolling campaigns mounted against the black empowerment aspects of Black Panther (Desto, 2018) and the female empowerment aspects of Captain Marvel (Zeitchik, 2019).

We collected data as follows. Fourteen seed webpages were chosen by internet search on the term “puppygate”; the dataset was extended by following links. Sites were labeled by hand according to which one of two sides in the debate they supported.

We will refer to the two sides as the “puppies” and the “others.” Links from the posts were followed, and text collected from the links, but only texts from news sites or those judged by a human annotator to be in the science fiction community (as fan, writer, or publisher), were included in the final dataset. Whenever possible multiple pages were collected from a site and

Table 2. Descriptions of three systems for which results are reported.

Link	This **baseline system** uses only community detection on the similarity graph that sums the co-reference similarity graph with the co-citation similarity graph.
Ling	This system uses the full pipeline described below, but only with proper names features. That is, no information from the links graph is used.
LingLink	This full pipeline system runs community detection on a graph that sums the similarity matrices from the co-reference similarity with the linguistic co-reference graph, LingCoRef in equation (1). The LingCoCit graph is not used in this study, because of the added challenge of resolving references to names as implicit references to blog sites.

concatenated together into a single site document. The sites were then hand-labeled as "puppy" or "other," depending on which side of the controversy they supported.

Methodology

We compare three systems in Table 2, Link, Ling, and LingLink; Link uses link-based information only (using hyperlinks to construct CoCit and CoRef); we will refer to the summed CoCit and CoRef graphs as CoCitRef. Ling uses linguistic information only (using proper names to construct LingCoRef); and LingLink combines CoCitRef and LingCoRef (as in equation (1)).

Ling and LingLink share the pipeline illustrated in Figure 7, which introduces three hyper-parameters whose effects are illustrated in the results discussed in the next section.

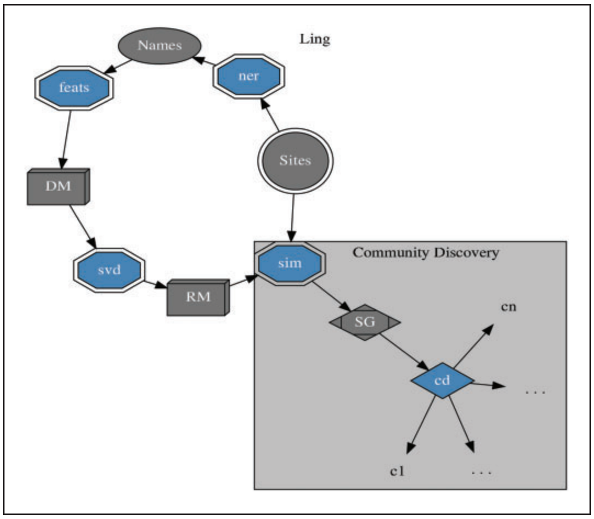

Figure 7. The community discovery system, distinguishing the language processing portion from community discovery proper. Beginning with Sites, the input text data, the Ling phase does named-entity recognition (ner), feature selection (feats), SVD-reduction (svd). The community discovery phase consists of construction of a similarity graph (sim) and community discovery proper (cd).

1. Names are extracted using the Stanford Named Entity Extractor (Finkel et al., 2005). Only person, names of works, and organization names are kept, with each site represented as a bag of names. Names are stemmed as described in the "Approach" section.
2. Feature weights are assigned using PMI. Features are filtered to the F most frequent. Thus, after feature reduction we have an N×F data matrix DM.
3. As happens often in high-dimensionality text-based document clustering, clustering performed directly on DM does not produce good results. We use singular value decomposition (SVD) to perform Dimensionality Reduction, to R, the number of reduced dimensions. Dimensionality Reduction maps from the N×F DM matrix to the N×R RM matrix.
4. We had our best success using the scikit-learn implementation (Pedregosa et al., 2011) of the stochastic SVD algorithm of Halko et al. (2011).
5. An N×N similarity matrix SM (LingCoRef) is constructed from the reduced data matrix RM using cosine as the similarity measure. In the case of the LingLink system, the similarity graph from the Link system, CoCitRef, is summed with LingCoRef. Although CoCitRef is already quite dense (see, e.g., Figure 6), LingCoRef, which includes linguistic similarity factors, is even denser, because language overlaps prevent 0-similarity pairs. Therefore, we use a pruned similarity graph SG for faction discovery. An edge exists between *i* and *j* in SG if either *j* is among the K most similar nodes to *i* or *i* is among the K most similar nodes to *j*, with the proviso that no edge can exist between vertices with similarity 0.
6. We run community detection using either Louvain or Newman on similarity graph SG.

Summarizing the system hyper-parameters: F is the number of features after filtering by frequency, R the number of reduced features after SVD, K the maximum number of neighbors in the similarity graph SG.

Results

Table 3 gives the results of our experiments for the Link, Ling, and LingLink systems using Newman and Louvain for community detection. Numbers shown are all AMI scores × 1000, and ± values show the 95% confidence half-intervals after 10 runs of each system. Indeterminacy was introduced by our stochastic SVD implementation, as well as by Newman, because of the Kernighan–Lin like refinement step the algorithm uses.

Table 3 shows that LingLink, the system combining Linguistic and link information, improves significantly on the baseline; note that all of the LingLink systems improve on their Ling counterparts, and none of the Ling systems improve on the baseline Link system, demonstrating fairly robustly that these linguistic features alone cannot converge on a solution of the community problem in an unsupervised setting. It is equally clear that the addition of the linguistic information adds value. The pattern of name usage does say something about the faction one belongs to.

Not shown in Table 3 is the comparison between two distinct systems that dispense with linguistic information, one that does community discovery on the links graph, and one that sums the CoCit and CoRef graphs. This is precisely the difference we looked at in our Polblogs example in Table 1, where the two methods produced virtually indistinguishable results; in this case, the links graph produced an AMI score of .142 and the CoCitRef graph produced an AMI score of .169. The higher of these was reported above.

Discussion

Our results show that adding linguistic information to the best Newman-algorithm based system gives a significant performance boost over the baseline system in reconstructing the communities in this fracture. Clearly the hyperlinks alone are not enough when they span between groups so often, and clearly there is some benefit in investigating the use of other kinds of similarity information.

Overall the performance of the faction detection system is promising, but still far short of satisfactory. Figure 8 provides a visualization of what the problem may be. The actual puppies/other classes are shown in the graph on the left and the Newman communities in the graph on the right. Note that the Newman communities agree much better with the grouping of the nodes chosen by the layout algorithm, suggesting that the Newman algorithm is doing a good job of finding a partition that respects the structure of the graph. The problem, then, may be that the similarity relations we have used to build that graph are not doing a good enough job of distinguishing the two factions.

Table 3. Upper table: AMI scores for the full pipeline systems with Newman. The baseline Link system achieved an AMI of .169 with a 95% confidence half-interval of 0, because the Newman algorithm found the same community on all runs. Lower table: same systems with Louvain.

			Newman	
K	F	R	Ling	LingLink
50	2K	3	118 ± 05	192 ± 07
		7	80 ± 17	189 ± 08
		25	101 ± 13	183 ± 45
	5K	3	127 ± 23	139 ± 33
		7	95 ± 12	194 ± 13
		25	92 ± 20	213 ± 40
75	2K	3	110 ± 13	**309 ± 31**
		7	95 ± 09	193 ± 08
		25	98 ± 09	188 ± 08
	5K	3	105 ± 25	098 ± 16
		7	99 ± 13	195 ± 18
		25	79 ± 20	188 ± 26

			Louvain	
K	F	R	Ling	LingLink
50	2K	3	117 ± 12	179 ± 13
		7	80 ± 06	164 ± 09
		25	57 ± 08	112 ± 17
	5K	3	83 ± 17	152 ± 15
		7	110 ± 13	154 ± 15
		25	54 ± 10	117 ± 16
75	2K	3	129 ± 11	179 ± 22
		7	89 ± 09	160 ± 05
		25	57 ± 10	129 ± 12
	5K	3	127 ± 22	161 ± 15
		7	104 ± 12	150 ± 15
		25	52 ± 14	136 ± 20

Note: Bold values signifies the best score.

Yet the information determining the factions is there, evident in the discussions in almost all the collected web pages, which almost all clearly contain language announcing the author's affiliation with one side and against the other. It is simply that the similarity features we have collected from the pages still contain too little information for the task.

One important limitation of this study is that we are not yet building a LingCoCit graph, tracking linguistic similarities among the sets of documents that refer to a document or author. Another is that the only language features used in this pilot study were proper names referring to persons, organizations, and works.

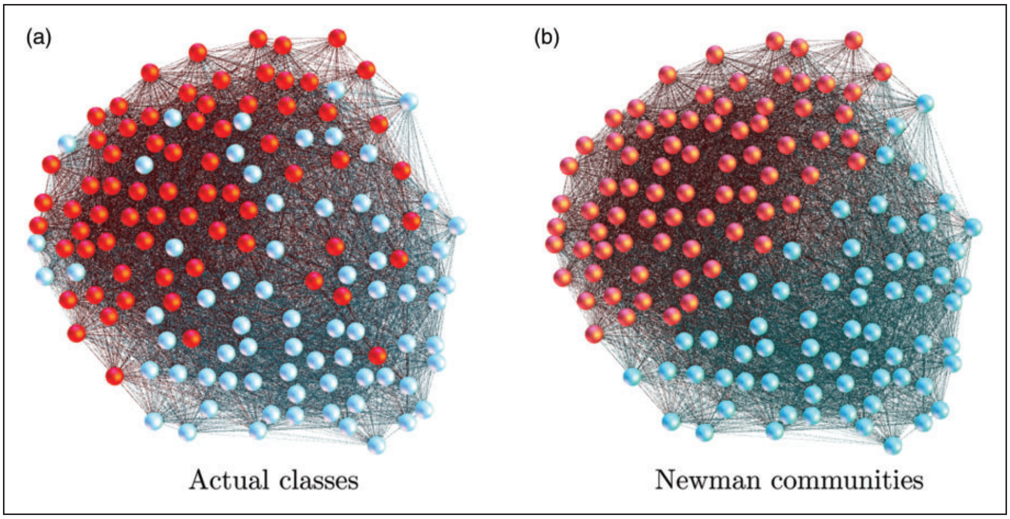

Figure 8. Depiction of the actual classes versus those discovered by the Newman algorithm. Note that the actual classes are not grouped in any single spatial region, suggesting that the graph may lack the information to recover the actual classes.

We know that proper names constitute only one of many linguistic strategies for referring to persons, organizations, and works. For example, there are full NP descriptions of people such as "the leader of the Social Justice Warriors" or of groups of works such as "every novel Arthur C. Clarke ever wrote," or linguistically complex titles referring to movies or books such as "An Inconvenient Truth." Each of these references is just as telling a marker of group identity as a use of a proper name such as "Al Gore" or "The New York Times." In addition, there is the rich range of language used for general cultural reference, such as Biblical citations ("John 3:17") or allusions to memes ("crazy cat videos"). Automatically recognizing each of these types of descriptions for what they are is a challenging Natural Language Processing task.

Conclusion

This study explores the use of linguistic information in community discovery, situating itself in a larger framework of similarity-based community detection. We have been at pains to argue two points: (a) that a similarity-based approach is not mutually exclusive with one that attends to the link-structure used in graph-based algorithms like Newman and Louvain; and (b) and that a similarity-based approach provides a flexible, principled way of combining different types of information in building the network. Hence its appropriateness for combining the very different kinds of information contained in link structure and referring language.

What we have shown in this very preliminary study is that, under certain conditions, adding linguistic information to link-based information can prove effective in identifying factions. Although a general characterization of when adding linguistic information helps is still an open problem, some features are clear. The existing link-based information must in some sense be inadequate for defining factions. One diagnostic of inadequacy seems to be that inter-faction links are relatively common. Another feature enhancing the utility of linguistic analysis is that reciprocal links be relatively rare. Both these properties distinguished our case study data from the Polblogs and karate club cases.

The significant improvement shown in this brief study, achieved using only a fragment of the information available in referring language, shows some of the promise of this extended form of faction discovery. In our discussion we suggested that the reason performance still fell short of satisfactory lay in the fact that the similarity graph still lacks the information to predict the fracture. That is, the place to look for the greatest improvement was not in better clustering or community detection algorithms, but in extracting more of the relevant information from the language. We hope though this pilot study to have provided some motivation for tackling the Natural Language Processing issues involved in a more complete analysis of referring language.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Jean M Gawron https://orcid.org/0000-0001-6288-4014

References

Adamic LA and Glance N (2005) The political blogosphere and the 2004 US election: Divided they blog. In: *Proceedings of the third international workshop on Link discovery*, pp.36–43. New York, NY: ACM Digital Press.

Aggarwal CC (2011) Introduction to social network data analytics. In: Aggarwal CC (ed.) *Social Network Data Analytics*. Boston, MA: Springer, pp. 1–15.

Aljaber B, Stokes N, Bailey J, et al. (2009) Document clustering of scientific texts using citation contexts. *Information Retrieval* 13: 101–131.

Aljaber B, Stokes N, Bailey J, et al. (2010) Document clustering of scientific texts using citation contexts. *Information Retrieval* 13: 101–131.

Anderson B (2003) *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. New York: Verso.

Biedenharn I (2015) Hugo awards fall victim to misogynistic and racist voting. *Entertainment Weekly*, 6 April.

Blondel VD, Guillaume J-L, Lambiotte R, et al. (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008: 10008.

boyd D and Crawford K (2012) Critical questions for big data. *Information, Communication & Society* 15: 662–679.

Bradshaw S (2003) Reference directed indexing: Redeeming relevance for subject search in citation indexes. In: *International conference on theory and practice of digital libraries*, pp.499–510.

Brooker P, Barnett J and Cribbin T (2016) Doing social media analytics. *Big Data & Society* 3(2): 1–12.

Chen C, Ibekwe-SanJuan F and Hou J (2010) The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology* 61: 1386–1409.

Christiaens T (2016) Digital subjectivation and financial markets: Criticizing social studies of finance with Lazzarato. *Big Data & Society* 3(2): 1–15.

Church KW and Hanks P (1990) Word association norms, mutual information, and lexicography. *Computational Linguistics* 16: 22–29.

Clauset A, Moore C and Newman ME (2008) Hierarchical structure and the prediction of missing links in networks. *Nature* 453: 98–101.

Clauset A, Newman ME and Moore C (2004) Finding community structure in very large networks. *Physical Review E* 70: 066111.

Colbert S (2014) The Colbert Report – Gamergate – Anita Sarkeesian. YouTube.

Correia L (2015) Sad puppies update: The nominees announced and why I refused my nomination. monster-hunternation.com, 4 April.

Cover TM and Thomas JA (2012) *Elements of Information Theory*. Oxford: John Wiley and Sons.

Cui W, Lin S, Tan L, et al. (2011) TextFlow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics* 17(2): 2412–2421.

Cutting DR, Karger DR, Pedersen JO, et al. (1992) Scatter/gather: A cluster-based approach to browsing large document collections. In: *Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval*, pp.318–329. New York, NY: ACM Digital Press.

de Solla Price DJ (1965) Networks of scientific papers. *Science* 149: 510–515.

Desto Y (2018) Rotten tomatoes is fighting back against white nationalist black panther trolls. Available at: https://www.vanityfair.com/hollywood/2018/02/rotten-tomatoes-black-panther-facebook-group (accessed June 7, 2016).

Dewan T and Squintani F (2016) In defense of factions. *American Journal of Political Science* 60: 860–881.

Finkel JR, Grenager T and Manning C (2005) Incorporating non-local information into information extraction systems by Gibbs sampling. In: *Proceedings of the 43nd annual meeting of the association for computational linguistics (ACL 2005)*, pp.363–370. Stroudsberg, PA: Association for Computational Linguistics.

Flood A (2014) Hugo award nominees withdraw amid 'Puppygate' storm. *The Guardian*, 17 April.

Fortunato S (2010) Community detection in graphs. *Physics Reports* 486: 75–174.

Garfield E (1955) Citation indexes for science: A new dimension in documentation through association of ideas. *Science* 122: 108–111.

Gawron, J. M., Gupta, D., Stephens, K., Tsou, M. H., Spitzberg, B., & An, L. (2012). Using group membership markers for group identification. In *Sixth International AAAI Conference on Weblogs and Social Media*. Menlo Park, CA: AAAI Press, pp. 467–470.

Girvan M and Newman ME (2002) Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99: 7821–7826.

Gui Q, Deng R, Xue P, et al. (2018) A community discovery algorithm based on boundary nodes and label propagation. *Pattern Recognition Letters* 109: 103–109.

Gupta D (2008) *Understanding Terrorism and Political Violence: The Life Cycle of Birth, Growth, Transformation, and Demise*. New York: Routledge.

Halko N, Tropp JA and Martinsson PG (2011) Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. *SIAM Review, Survey and Review Section* 53(2): 217–288.

Haveliwala TH, Gionis A, Klein D, et al. (2002) Evaluating strategies for similarity search on the web. In: *Proceedings of the 11th international conference on World Wide Web*, pp.432–442. New York, NY: ACM Digital Press.

Hunter JD (1992) *Culture Wars: The Struggle to Define America*. New York: Basic Books.

Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* 46: 604–632.

Manning CD, Raghavan P and Schütze H (2008) *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Martin GRR (2015) Puppygate. grrm.livejournal.com.

Newman MEJ (2006a) Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103: 8577–8582.

Newman ME (2006b) Finding community structure in networks using the eigenvectors of matrices. *Physical Review E* 74: 036104.

Nobyline (2015) Hugo award nominations spark criticism over diversity in sci-fi. *The Telegraph*, 8 April.

Papakyriakopoulos O, Hegelich S, Shahrezaye M, et al. (2018) Social media and microtargeting: Political data processing and the consequences for Germany. *Big Data & Society* 5(2): 1–15.

Parkin S (2014a) Gamergae: A scandal erupts in the videogame community. *The New Yorker*, Oct 17, 2014, 1–4.

Parkin S (2014b) Zoe Quinn's depression quest. *The New Yorker*, Sep. 9, 2014.

Pedregosa F, Varoquaux G, Gramfort A, et al. (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Persico N, Pueblita JC and Silverman D (2011) Factions and political competition. *Journal of Political Economy* 119: 242–288.

Pons P and Latapy M (2005) Computing communities in large networks using random walks. *International symposium on computer and information sciences*. Heidelberg: Springer, pp. 284–293.

Qiu J and Lin Z (2014) D-HOCS: An algorithm for discovering the hierarchical overlapping community structure of a social network. *Journal of Intelligent Information Systems* 42: 353–370.

Raghavan P (2014) It's time to scale the science in the social sciences. *Big Data & Society* 1(1): 1–4.

Raghavan UN, Albert R and Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E* 76(3): 036106.

Rapoport M (2015) The culture wars invade science fiction. *The Wall Street Journal*, 15 May.

Ritchie A, Robertson S and Teufel S (2008) Comparing citation contexts for information retrieval. In: *Proceedings of the 17th ACM conference on information and knowledge management*, pp.213–222. New York, NY: ACM Digital Press.

Robertson S, Zaragoza H and Taylor M (2004) Simple BM25 extension to multiple weighted fields. In: *Proceedings of the thirteenth ACM international conference on information and knowledge management*, pp.42–49. New York, NY: ACM Digital Press.

Rosvall M, Axelsson D and Bergstrom CT (2009) The map equation. *The European Physical Journal Special Topics* 178(1): 13–23.

Salton G (1971) *The SMART Retrieval System*. Upper Saddle River, NJ: Prentice-Hall.

Salton G and McGill MJ (1983) *Introduction to Modern Information Retrieval*. New York: McGraw Hill.

Sasahara K, Hirata Y, Toyoda M, et al. (2013) Quantifying collective attention from tweet stream. *PLoS ONE* 8(4): e61823.

Scalzi J (2015) A note about the Hugo nominations this year. whatever.scalzi.com, 6 April.

Segev E, Nissenbaum A, Stolero N, et al. (2015) Families and networks of internet memes: The relationship between cohesiveness, uniqueness, and quiddity concreteness. *Journal of Computer-Mediated Communication* 20(4): 417–433.

Small H (1973) Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science* 24: 265–269.

Tang J, Sun J, Wang C, et al. (2009) Social influence analysis in large-scale networks. In: *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining*, June, pp.807–816. New York: ACM.

Tumasjan A, Sprenger TO, Sandner PG, et al. (2010) Predicting elections with twitter: What 140 characters reveal about political sentiment. In: *Fourth international AAAI conference on weblogs and social media*, Washington, DC, USA, 23–26 May 2010, pp.178–185. Menlo Park, CA: The AAAI Press.

Vinh NX, Epps J and Bailey J (2010) Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research* 11: 2837–2854.

Wang Y and Kitsuregawa M (2004) Enhancing contents_link coupled web page clustering and its evaluation. In: *Proceedings of data engineering workshop (DEWS 2004)*, pp.499–506. New York, NY: ACM Digital Press.

Weiss R, Vélez B and Sheldon MA (1996) Hypursuit: A hierarchical network search engine that exploits content-link hypertext clustering. In: *Proceedings of the seventh ACM conference on hypertext*, pp.180–193. New York, NY: ACM Digital Press.

White HD and McCain KW (1998) Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *Journal of the American Society for Information Science* 49: 327–355.

Wingfield N (2014) Intel pulls ads from site after 'Gamergate'. *New York Times*, 2 October.

Zachary WW (1977) An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* 33: 452–473.

Zeitchik S (2019) Captain Marvel: How the trolls always win – Until they don't. *Washington Post*, 7 March.

Zhao D and Strotmann A (2014) The knowledge base and research front of information science 2006–2010: An author cocitation and bibliographic coupling analysis. *Journal of the Association for Information Science and Technology* 65: 995–1006.