Network Science **1** (2): 170–190, 2013.
doi:10.1017/nws.2013.12

Characterizing ego-networks using motifs

PÁDRAIG CUNNINGHAM, MARTIN HARRIGAN, GUANGYU WU and DEREK O'CALLAGHAN

School of Computer Science and Informatics, University College Dublin, Ireland
(*e-mail:* padraig.cunningham@ucd.ie)

Abstract

We assess the potential of network motif profiles to characterize ego-networks in much the same way that a bag-of-words strategy allows text documents to be compared in a vector space framework. This is potentially valuable as a generic strategy for comparing nodes in a network in terms of the network structure in which they are embedded. In this paper, we consider the computational challenges and model selection decisions involved in network motif profiling. We also present three case studies concerning the analysis of Wikipedia edit networks, YouTube spam campaigns, and peer-to-peer lending in the Prosper marketplace.

1 Introduction

Networks model objects and the relationships between those objects. Social network analysis and network data analysis more generally are based on the fundamental assumption that the network structure around an object can tell us a lot about that object. This is a well-established idea not only in social network analysis (Borgatti & Everett, 1992; Krause et al., 2007; Juszczyszyn et al., 2008) but also in bioinformatics (Milo et al., 2004; Vazquez et al., 2004; Pržulj, 2007), in spam filtering (Boykin & Roychowdhury, 2005; Becchetti et al., 2008; Kamaliha et al., 2008), and in telecommunications (Allan et al., 2009).

In this paper, we consider the challenge of assessing the similarity between *ego*-networks, that is the local network structure around specific objects of interest. This similarity assessment is important because in many situations similar objects will be embedded in similar networks—indeed in many situations the similarity in network structure may be the main visible evidence of similarity.

Graph similarity is one of the core challenges in learning and analysis tasks associated with networks. It is computationally challenging because subgraph isomorphism is NP-complete. Even if it were a tractable solution, subgraph isomorphism does not solve the problem because graphs can be *similar* while having little structure that matches exactly. This is true in the same way that two text documents can be similar without matching at a sentence level.

The solution to the problem of graph similarity that we explore here is analogous to the vector space model for document similarity in information retrieval. In information retrieval, documents are represented as feature vectors of index terms that map each document to a point in a high-dimension space. Here, we propose to map each ego-network as a feature vector of motif counts so that networks can be compared based on the similarity of these feature vectors (see Figure 1). This allows

us to compare Wikipedia pages in terms of the similarity of their edit networks, or identify YouTube users that are similar to known spammers or cluster users in a peer-to-peer lending system in terms of their lending behavior.

An important principle in this strategy is that all motifs up to a given size (typically five nodes) are used in the profiling. There is no *a priori* selection of motifs that are considered to be important for the particular classification task. The principle is that the full set of motifs should carry *sufficient statistics* to characterize the ego-network. If the motif-based representation is to be used in a supervised learning setting, there is always the potential to use feature selection methods to identify the discriminating motifs. This in turn offers insight into the nature of the classification problem.

The paper proceeds with a simple example in the next section to illustrate the idea of profiles based on motif counts. Then, before presenting a series of case studies we provide a review of related research in Section 3. The most detailed example in the paper is the example in Section 4 illustrating the relationship between edit network structure and article quality in Wikipedia. There follows two further examples on YouTube spam in Section 5 and on Prosper in Section 6. The merits and shortcomings of motif profiling are discussed in the conclusions in Section 7.

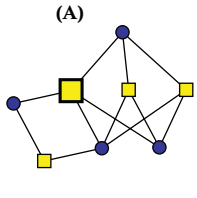
2 A simple example

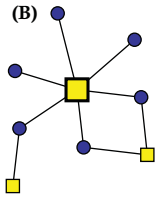
The objectives of motif-based characterization are best illustrated with a simple example. Figure 1 shows three simple Wikipedia edit networks of the type discussed in Section 4. These are the *target* networks where we count instances of the motifs—we can also refer to the motifs as *query* networks. Networks A and C have some of the characteristics of high-quality networks in that editors collaborating on the *ego*-page have collaborated on other pages, while the B network is typical of a "Start" class article. An examination of the motif counts in the table shows that the motif counts do discriminate between the high- and low-quality networks. Network B has high counts of the star motifs 4 and 8 that indicate multiple editors collaborating on a page. However, motifs 13, 14, and 15 that indicate dense collaboration across multiple articles are absent in network B.

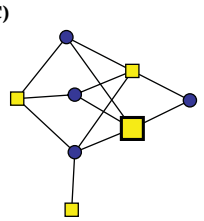
There are some technical details in the profiling process that are worth highlighting. The profile covers *all* motifs of up to five nodes that can occur in a bipartite network of editors and articles. The motif counts cover all occurrences of motifs in the network. In some circumstances, it may be valid to limit the count to motifs occurrences that actually include the ego node—for instance, there are examples of motif 3 in networks A and C that do not contain the ego.

Another important issue is the question of *induced* motif counts. An induced motif must contain all edges between its nodes that are present in the target network, whereas a non-induced motif need not. To illustrate this, we can take motif 7 as the target network and motif 6 as the query motif. There are two occurrences of motif 6 in motif 7 if we count without the "induced" constraint. However, if we consider induced motifs then the count is zero. A clear-cut example of where this arises in Figure 1 is in the count of motif 17 in network C. While there are examples of motif 17 in C there are no induced examples. The appropriateness of induced

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press







	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	12	13	13	6	6	16	10	1	1	4	4	8	8	3	3	4	4
B	9	16	3	20	0	13	1	15	0	22	0	4	0	0	0	2	0
C	12	13	11	8	5	15	8	2	1	10	4	4	9	4	1	0	0

Fig. 1. An example of motif counts in three ego-networks A, B, and C. The networks are simple versions of the Wikipedia edit networks discussed in Section 4. The light nodes are articles and the dark nodes are editors, the large square indicates the ego node. (color online)

versus non-induced counting depends on the circumstances. Induced motif counting is appropriate where the *absence* of an edge is significant.

3 Related research

In this paper, we bring together research on network motif analysis and egocentric network analysis so we review the relevant research in subsections under these headings. We also provide overviews of research on exponential random graph ($p*$) models and relevant research on graph kernels.

3.1 Network motif analysis

It is probably reasonable to identify the publication in Science in 2002 of the work by Milo et al. (2002) as a key milestone for network motif analysis in computational science. However, the roots of network motif analysis in social science are a good deal older. While graph theoretic analysis has a long history in social science, for our purpose it is the social science research that emphasizes local rather than global network structure that is important (Moreno, 1934; Davis, 1963; Holland & Leinhardt, 1976).

In this respect, the analysis by Moreno (1934) of sociograms in terms of chains, triads, and stars could be considered the genesis of network motif analysis. The work by Holland and Leinhardt (1976), which emphasizes the insight offered by local rather than global analysis in social graphs, is also an important milestone. Holland and Leinhardt introduce the notion of a triad census which is a count of the occurrences of 16 possible triads in a graph (see Figure 2). They propose that, in a graph with one node type and directed edges, the counts of the 16 possible triads present a useful set of summary statistics for the graph.

Milo et al. (2002) use triad counts as summary statistics in their analysis and also consider 4-node motifs. Because the objective of their work is to discover motifs that occur more frequently than would be expected they also perform motif

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

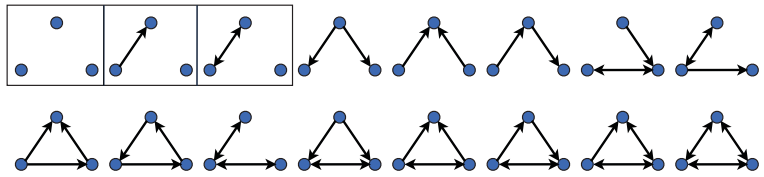


Fig. 2. The 16 triads used in the analysis by Holland and Leinhardt (1976). Milo et al. (2002) also use these triads to produce summary statistics; however, they do not consider the first three that are not completely connected. (color online)

counting on randomized versions of the networks. The networks are randomized by rewiring at random while preserving the degree of individual nodes. Motif counts in these rewired networks provide a baseline or null model against which frequently occurring motifs can be identified.

This strategy whereby motifs are considered interesting if they occur significantly more often in the real data than in a random model has been the subject of some criticism. Artzy-Randrup et al. (2004) have pointed out that in many cases the random models used are not appropriate and can suggest significance when the evidence is not really there. This null model problem is not an issue when comparing ego-networks as the motif statistics are used directly to compare ego-networks without reference to a null model.

Because the vocabulary of motifs increases very rapidly as more nodes are included, it is important to consider redundancy across motif counts. Clearly, counts of motifs of the same size may be highly correlated and counts of higher order motifs may not contain additional information that is not contained in their constituent motifs. In this regard, Faust (2007) has shown that a triad census contains little information that is not already available from a dyad census plus some other simple statistics. She compares the triad census of 82 networks against lower-order characterizations based on network density, out- and in-degree distribution, and the dyad census. She shows that 90% of the triadic structure is accounted for by these lower-order properties. In a sense, this is reassuring as it suggests that it is not necessary to count higher-order motifs in order to characterize local network structure. We pick up this issue again in Section 4.1.

Kalish & Robins (2006) have analyzed the relationship between psychological predispositions and local network structure. They examine how triad counts in one-step ego-networks entailing three types of edges (strong, weak, and no tie) predict psychological traits such as neuroticism, extroversion, and group or individual focus. With three types of edges, their triad census has nine types of triads. While this is a relatively small motif vocabulary, it nevertheless reduces to just three principal components that account for 72% of the variance in the data.

This is a clear demonstration that motif counts can be highly correlated so there is considerable scope for effective dimension reduction. However, we need to remember that features that do not capture much variance in the data may be quite discriminating between classes. This is the whole motivation behind discrimination-based dimension reduction techniques such as linear discriminant

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

analysis (Fisher, 1936) that focus on discrimination rather than data distribution. This issue of dimension reduction and classification is addressed in Section 4.3.1.

Beyond the original work in social science and bioinformatics, network motif analysis has also found application in spam detection. Within a network built from email addresses (Boykin & Roychowdhury, 2005), a low clustering coefficient (based on the number of triangle structures within a network) may indicate the presence of spam addresses, with regular addresses generally forming close-knit communities, i.e. a relatively higher number of triangles. Becchetti et al. (2008) made use of the number of triangles and clustering coefficient as features in the detection of web spam. These two features were found to rank highly within an overall feature set. Motifs of size three (triads) have also been used to detect spam comments in networks generated from blog interaction (Kamaliha et al., 2008). It was found that certain motifs were likely to indicate the presence of spam, based on comparison with corresponding random network ensembles.

In telecommunications, network motifs have also been used to characterize network traffic (Allan et al., 2009). A network was created for each application (e.g. HTTP), and nodes within the network were classified using corresponding motif profiles.

3.2 Egocentric networks

An egocentric network is the local social network around an individual. It will include all individuals directly connected to the *ego*, the central node in the network. It may also include the neighbors of the ego's neighbors, that is individuals two hops from the ego. If the network structure around an individual can tell us something about that individual, then this information should be contained in the ego-network.

Egocentric analysis has a long history within the field of social network analysis (White et al., 1976; Borgatti & Everett, 1989; Wellman, 1993). Some representative recent work (Lubbers et al., 2010) describes a dynamic egocentric network analysis of Argentinean immigrants in Spain. The analysis comprised qualitative interviews and a quantitative analysis at three distinct levels (ego-alter dyads, alter-alter dyads, and networks). The quantitative analysis investigated the characteristics of the egos, the structural characteristics of the networks, the characteristics of the ego-alter dyads and alters, the structural positions of the alters, and the characteristics of the alter-alter dyads. The composition of the egocentric networks was visualized using clustered networks (Brandes et al., 2008), where the size of four nodes encodes the number of people in each of four groups (origin, fellows, host, and transnationals) and the thickness of the edges quantifies the amount of communication between the groups.

Welser et al. (2007) present an analysis of roles in an online discussion group. They visualize egocentric networks and posting habits. Through visual inspection, they identify three types of posters: answer people, discussion people, and disruptors. Similarly, Stoica and Prieur (2009) analyze egocentric networks in a large mobile phone call network. They partition the nodes according to roles and validate their results using network attributes. Antiqueira and da Fontoura Costa (2009) present a methodology for analyzing non-overlapping subnetworks, their interrelationships, and their distribution in a network. They analyze four random network models

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

Table 1. *The configurations used in the analysis of biological networks by Saul and Filkov et al. (2007).*

Configuration	Description
k-star	The number of nodes in the network with exactly k adjacent edges with unconnected end points.
k-cycle	The number of k-cycles in the network.
k-degree	The number of nodes in the graph with degree k.
k-edgewise shared partners	The number of edges in the network connecting nodes that have exactly k shared partners.
Geometrically weighted degree	The weighted sum of the counts of each degree, weighted by the geometric sequence $(1 - e^{-\alpha})^i$ where α is a decay parameter.
Geometrically edgewise shared partners	The weighted sum of the number of edges connecting nodes having exactly i shared partners weighted by the geometric sequence $(1 - e^{-\alpha})^i$.
Isolates	The number of nodes in the network with no neighbors.

and five real-world networks and show, for example, that the real-world networks have similarities with combinations of the random network models. These three analyses (Welser et al., 2007; Stoica & Prieur, 2009; Antiqueira & da Fontoura Costa, 2009) choose a number of network statistics, for example, degree, clustering coefficient, and local triangle count, when characterizing subnetworks. The choice is often specific to the task at hand. In adapting network motif analysis to egocentric networks, we are not committing to a particular choice of network statistics.

Clearly, comparisons and the determination of similarity is a core issue in an egocentric network analysis. Using motif counts to determine similarity in ego-networks avoids the null model problem associated with network motif analysis identified by Artzy-Randrup et al. (2004) since comparisons within collections of ego-networks do not require a null model.

3.3 Exponential random graph (p*) *models*

Exponential random graph models (ERGMs) can be seen as a development on the basic triadic model presented by Holland and Leinhardt (1976) (see Figure 2). A network of n nodes can be represented by an $n \times n$ matrix $\mathbf{X}$ where $\mathbf{X}_{i,j} = 1$ if there is an edge between i and j. ERGMs present a descriptive model that provides a probability density function $P(\mathbf{X} = \mathbf{x})$ that allows us to calculate the probability of $\mathbf{x}$, an observed instance of $\mathbf{X}$. The model is based on a set of explanatory variables $\mathbf{z}(\mathbf{x}) = z_1(\mathbf{x}), z_2(\mathbf{x}), \ldots, z_r(\mathbf{x}))$. These explanatory variables (also called configurations or local patterns) can be any graph or node statistic, some examples are listed in Table 1.

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

$P(\mathbf{X} = \mathbf{x})$ can be modeled with a log linear model as follows.

$$log(P(\mathbf{X} = \mathbf{x})) \propto \theta_1 z_1(\mathbf{x}) + \theta_2 z_2(\mathbf{x}) + \ldots + \theta_r z_r(\mathbf{x})$$
$$\propto \theta' \mathbf{z}(\mathbf{x})$$

This allows us to estimate the probability of an observed network $\mathbf{x}$:

$$P(\mathbf{X} = \mathbf{x}) = \frac{e^{\theta' z(x)}}{\kappa(\theta)} \tag{1}$$

based on a vector of parameters θ, where $\kappa(\theta)$ is a normalizing term (Anderson et al., 1999). The identification of appropriate explanatory variables $\mathbf{z}(\mathbf{x})$ and the estimation of the model parameters θ have received much attention in the social networks research literature (Robins et al., 2007a; Robins et al., 2007b).

Saul and Filkov et al. (2007) present an analysis of biological networks using ERGMs that has similar objectives to the work presented here. They show that what they call a topological profile based on ERGMs is effective for clustering biological networks based on function. In their analysis, they use statistics based on the features shown in Table 1.

We see our method as closely related to this ERGM strategy. The main difference is that the ERGM strategy has a formal foundation where the objective is to produce a descriptive model of the network. Then the coefficients of this model can be used as a feature set that characterizes the network rather than the original network statistics. It is an open question whether these model parameters are more effective than the original network statistics for classification and clustering—which is our objective here.

3.4 Graph kernels

Since the objective with motif analysis is to produce a profile that characterizes a network, it is worth also looking at graph kernels that directly compare two networks. Formally speaking, a graph kernel is a function that quantifies the similarity between two graphs G_i and G_j

$$k(G_i, G_j) = \langle \phi(G_i), \phi(G_j) \rangle \tag{2}$$

While $\phi(G)$ is a function that maps the graph into a vector space where a dot product is performed, there may be no direct representation of $\phi(G)$. Luo et al. (2003) present a fairly fundamental family of graph kernels based on graph spectra. If X_i is the adjacency matrix of graph G_i then the spectral decomposition of X_i is

$$X_i = V_i \Lambda_i V_i^T \tag{3}$$

where V_i is a matrix constructed from the eigenvectors of X_i and Λ_i is a diagonal matrix of the eigenvalues of X_i. A basic characterization of G_i that can be derived from this spectral decomposition is the vector of k leading eigenvalues. Luo et al. show how basic subgraph features such as volume (sum of the degrees of nodes belonging to the subgraph), perimeter, length (number of edges exiting the subgraph), and Cheeger constant (the extent to which there is a bottleneck in the external connections from the subgraph) relate to the spectral decomposition.

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

Table 2. *Datasets analyzed.*

	Good		Medium	
Datasets	Classes	Count	Classes	Count
History-F-S	F	149	S	299
US-F-S	F	272	S	300
Meteorology-F-S	F	131	S	300
History-FG-CS	FG	440	CS	588
US-FG-CS	FG	565	CS	598
Meteorology-FG-CS	FG	431	CS	600

This basic idea has been extended in a number of ways to produce graph kernels that capture notions of similarity focused on a variety of different criteria (Shervashidze et al., 2009), for example random walks (Gärtner et al., 2003; Kashima et al., 2004), shortest paths (Borgwardt & Kriegel, 2005), subtrees (Ramon & Gärtner, 2003), and cycles (Horváth et al., 2004). In turn, Shervashidze et al. (2009) have proposed graphlet kernels that compare graphs in terms of the distribution of motifs which they term graphlets. This is closely related to the motif-based characterization presented here except that the graphlets (i.e. motifs) are hidden in the kernel implementation. Shervashidze et al. have focused on efficient kernel computation at the expense of interpretability. In the next section, we demonstrate the benefits of interpretability whereby the identification of discriminating motifs provides an insight into the nature of the networks.

4 Case study on Wikipedia

In Section 2, we introduced Wikipedia edit networks (Figure 1) as an example of ego-networks that might be suitable for motif-based analysis. We develop this example in this section focusing on the objectives of classification and anomaly detection. Before looking at performance against these objectives we consider some important implementation issues—particularly the impact of motif size and also the need for normalization.

The experiments are based on the articles from three collections on *History*, *United States*, and *Meteorology*. These collections were selected because they include a large number of articles, especially a sufficient number of *featured articles*. The evaluation considered Wikipedia articles from four different quality classes, featured articles (F), good article (G), C-class articles (C), and start articles (S). We consider the first two classes to be articles of high quality while the last two are of medium or low quality.

From these collections we created six datasets (see Table 2) representing two types of classification challenge, an easy challenge comparing F and S articles and a harder challenge with F and G as the good quality class and C and S as the low/medium class.

To build our edit networks, we applied some rules to filter the raw data. Firstly, as some articles have a long edit history, we only considered the last 200 revisions of the articles, and extracted the editors who made these revisions. We also considered all articles that are connected by hyperlinks from the originating or ego articles. We

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

Table 3. *Motif counting times for motifs sets of different sizes on a test collection of 600 networks.*

	No. of motifs	Counting time
Up to 2 nodes	1	<1 minute
Up to 3 nodes	3	1 minute
Up to 4 nodes	7	11 minutes
Up to 5 nodes	17	251 minutes
Up to 6 nodes	44	8247 minutes

retained the linked-to articles that have been edited by at least one of the ego article editors.

4.1 Impact of motif sizes

Faust (2007) considers motifs of size two, Holland and Leinhardt (1976) work with motifs of size three, and Milo et al. (2002) consider motifs of size three and four. In Figure 1, we present all motifs up to five nodes. Clearly, some larger motifs will have the potential to capture some special characteristics of the networks. On the other hand, motifs are compositional so there will be diminishing returns in terms of characterization power as more motifs are added. There is also the problem that computation cost increases steeply as motifs get larger. In Figure 3 and Table 3, we present a simple analysis to illustrate this tradeoff.

With the objective of assessing the impact of motif size on classification accuracy and motif counting time, we selected 100 featured and 100 start articles from each of the *History*, *United States*, and *Meteorology* datasets. We performed motif counting to produce characterizations considering motifs from two nodes to six nodes in size. The `FANMOD` software (Wernicke & Rasche, 2006) was used to do the motif counting—it is designed to output all induced subgraph instances for a particular size in a given target network.

For two nodes, there is only one motif so the characterization is simply an edge count. At five nodes, the motif set is the 17-node set shown in Figure 1. The 3-node and 4-node collections are respectively the first three and the first seven motifs in Figure 1. There are 44 motifs in the 6-node collection.

The increase in processing time for the larger motifs can be seen in Table 3. The 4-node set takes 11 minutes, the 5-node set takes four hours, and the 6-node set takes five days. On the positive side, the diminishing returns for the larger motif sizes is evident in the classification accuracies shown in Figure 3. This shows average cross-validation accuracies for logistic regression and random forest classifiers (Cunningham et al., 2008) for the different motif sizes (see Section 4.3 for more detail on the supervised learning methods used in this evaluation). Clearly, the huge increase in computation to do the 6-node counting returns little in classification accuracy. For this reason, for the remainder of this paper, we consider motifs up to just five nodes.

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

Table 4. *Classification accuracies comparing raw and normalized data using cross-validation on a balanced dataset of 200 samples. Entries in bold indicate the best results.*

	Logistic		Random forest (100 trees)	
Datasets	Unnormalized	Normalized	Unnormalized	Normalized
History	**80.0%**	79.5%	78.5%	**80.5%**
USA	82.5%	**85.5%**	**82.0%**	80.0%
Meteorology	82.0%	**87.0%**	79.5%	**81.5%**
HUM	83.2%	**84.3%**	**82.3%**	80.8%

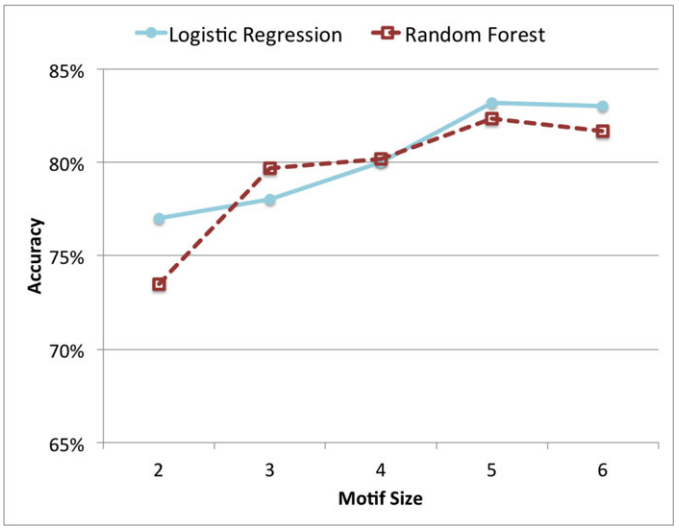


Fig. 3. Classification accuracies for different sizes of motif. The number of motifs at each size is shown in Table 3. These results are obtained using cross-validation on a balanced dataset of 600 examples. (color online)

4.2 Impact of normalization

The results for 2-node motifs in Figure 3 show that a simple edge count has significant classification power. This is because the edit networks for featured articles are inclined to be larger than those for start articles. However, start articles can have large networks and featured articles have small networks in some cases. Because of this it is worth assessing the impact of normalizing the motif counts to remove the effect of motif size. Table 4 shows the impact of L2 normalization (feature vectors normalized to unit length) on four datasets. *HUM* is the three individual datasets collected into a large dataset of 899 examples. While normalization does not have a clear impact on classification accuracy, it is probably a good idea to normalize the data in order to get rid of the size bias—a bias that is helpful with these data.

4.3 Classification

In this subsection, we look in more detail at classification accuracy using counts of motifs of up to five nodes as shown in Figure 1. We assess the performance of logistic regression, random forests, k-nearest neighbor (k-NN), and support vector machines (SVM) on the datasets summarized in Table 2.

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

Table 5. *Cross-validation results on the complete datasets (F versus S). Entries in bold indicate the best results.*

	Logistic		Random forest		*k*-NN		SVM	
Datasets	Acc.	ROC	Acc.	ROC	Acc.	ROC	Acc.	ROC
History-F-S	80.4%	0.88	**82.6%**	0.89	76.8%	0.73	79.5%	0.73
US-F-S	**87.2%**	0.94	85.8%	0.93	85.0%	0.85	83.4%	0.83
Meteorology-F-S	**81.0%**	0.87	78.9%	0.88	75.9%	0.72	75.2%	0.61

Table 6. *Cross-validation results on the harder datasets (FG versus CS). Entries in bold indicate the best results.*

	Logistic		Random forest	
Datasets	Acc.	ROC	Acc.	ROC
History-FG-CS	**65.7%**	0.70	65.3%	0.70
USA-FG-CS	**71.8%**	0.79	70.8%	0.79
Meteo-FG-CS	60.9%	0.67	**66.4%**	0.72

First, we applied the four classifiers on the three complete simple datasets (F versus S). The results in terms of classification accuracy and ROC area are presented in Table 5. The best classification accuracy is achieved with logistic regression scoring above 80% in all cases.

A clear pattern in these results is that random forest and logistic regression are performing better than *k*-NN and SVM. This pattern was maintained in our other evaluations so, to simplify the picture, we do not report further results using *k*-NN or SVM. The good performance of logistic regression is perhaps not surprising given that we are operating in a 17-dimension space where linear separability is easy to achieve and a linear classifier can produce reasonable accuracy. It is not surprising that random forest performs well as it is an ensemble method.

In the next evaluation, we tackle the larger datasets where high-quality articles include both featured and good articles and lower-quality articles are both C class and start articles (Table 6). Each of the datasets contains over 1,000 articles (see Table 2) and the classification is more difficult because the distinction between the two classes is less clear. Accuracy falls as expected but roughly two-thirds of articles are still classified correctly.

4.3.1 Feature selection

Given our objective of identifying the useful *signal* in edit network motifs for predicting Wikipedia article quality, we now turn to the correlations and contributions of individual motifs. In Section 3.1, we briefly reviewed other work (Kalish & Robins, 2006; Faust, 2007) that showed redundancy due to correlations in motif counts. If we can identify a small selection of motifs that have the classification power of the full set of motifs, then this tells us which motifs are characteristic of good-quality collaboration. This will also allow us to simplify the motif characterization process.

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

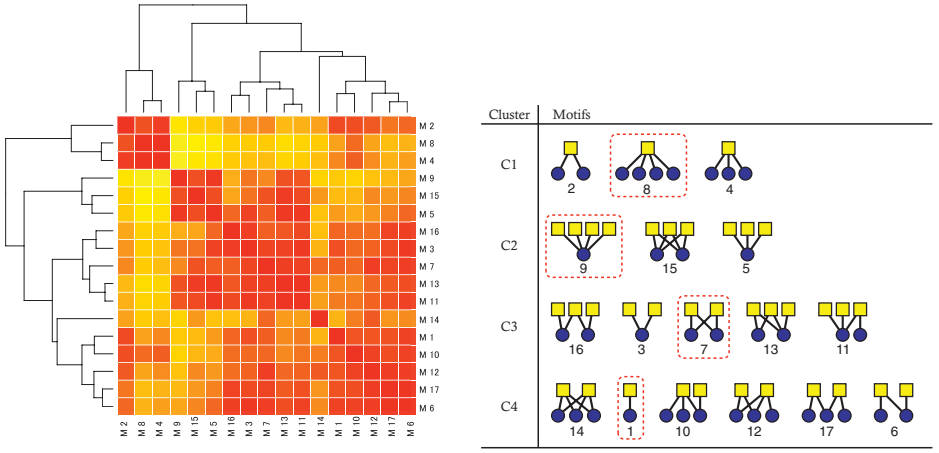


Fig. 4. A hierarchical clustering of network motifs with clustering based on correlations in motif counts. Darker colors in the heatmap indicate high correlations between motif counts, e.g. counts of M2, M4, and M8 are highly correlated. The dendrogram shows four clear clusters and the corresponding motifs are shown in the table on the right. (color online)

Our objective here is to cluster motifs based on correlated counts and then select a subset of motifs that are representative of the clusters. When we do this with the History data, we get the hierarchical heatmap shown in Figure 4. If we split the motif set into four clusters according to the obvious subtrees in the hierarchy, the four motif clusters are {M2, M8, M4}, {M9, M15, M5}, {M16, M3, M7, M13, M11}, and {M14, M1, M10, M12, M17, M6}.

Next, we would like to select representatives from each of these clusters that are easier to count. While the general problem of network motif counting is computationally expensive, there are certain motifs such as stars and cycles (Paton, 1969) that are easier to count. It transpires that we can select motifs M1, M7, M8, and M9 as representatives of each of the four clusters that satisfy this criterion (highlighted in Figure 4).

Classification performance using just this set of four motifs is shown in Table 7. There is a significant fall-off with logistic regression with reductions of about 2.5% on the USA and Meteorology data. Performance holds up better using the random forest classifier with performance on the three datasets using just four motifs within 1% of that using all motifs. If M7 is dropped from this group of four then accuracy drops by a further 2% to 4% indicating that the information it contains about collaboration is useful for classification.

Perhaps it is not surprising that the random forest classifier performs better than logistic regression in the lower-dimension space given that logistic regression requires linear separability.

5 YouTube spam

In our second case study, we consider the challenge of identifying *spammers* operating on YouTube. Spamming is possible on YouTube through the facility that permits users to post comments to videos. Orchestrated campaigns target popular YouTube videos with bot-posted spam comments. These campaigns are often of a recurring

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

Table 7. *Classification results for using only four motifs.*

	Logistic		Random forest	
Datasets	All motifs	4 motifs	All motifs	4 motifs
History-F-S	80.4%	82.4%	82.6%	81.7%
USA-F-S	87.2%	84.8%	85.8%	85.0%
Meteo-F-S	81.0%	78.7%	78.9%	78.7%

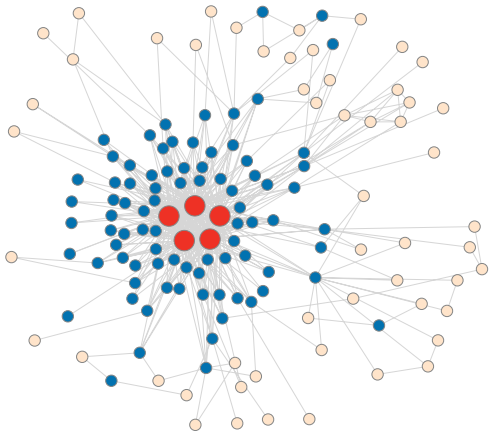


Fig. 5. Spam campaign targeting YouTube, with a small number of accounts each commenting on many videos. Darker nodes are videos, lighter nodes are user accounts, the five central spam accounts are highlighted. (color online)

nature, operating with periodic bursts (Xie et al., 2008; Gao et al., 2010) of activity on a continual basis, with different campaign strategies being employed. For example, a real network associated with a typical campaign is shown in Figure 5, where the strategy is to use a small number of spam accounts to target a large number of videos. The objective in this case study is to see if such spam accounts have characteristic network motif profiles.

Because we are interested in network motif profiling as a generic network analysis mechanism, our strategy has been to consider *all* motifs. We do not assume any prior knowledge of the nature of spam campaigns and motifs that might be characteristic of spammers. However, the enumeration of all motif instances present in the user networks can be a lengthy process, so the case study includes some motif subset selection as in the Wikipedia case study. We have found that certain discriminating motifs may be used to identify particular strategies and the associated users as they periodically recur.

5.1 YouTube comment networks

A YouTube dataset was collected for the purpose of investigating contemporary recurring spam campaigns (O'Callaghan et al., 2012). We periodically retrieved the details and comments of popular videos on a continual basis, using the *Most Viewed*

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

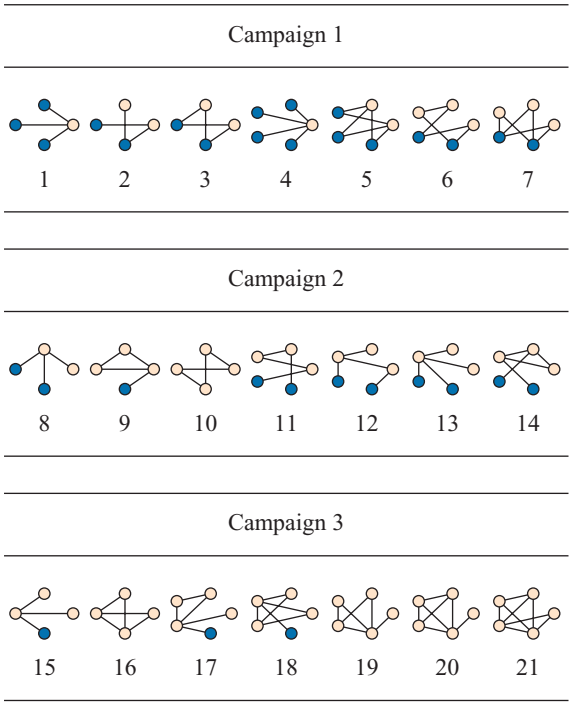


Fig. 6. Selected discriminating YouTube motifs—darker nodes are videos, lighter nodes are users. (color online)

standard feed provided by the YouTube Data API. We focus on the more popular videos because they have a higher probability of attracting attention from spammers because they ensure a larger audience.

Our methodology requires the generation of a network to represent the comment posting activity of users to a set of videos. Comments made during a specified time interval are selected from the dataset, which are then converted to a set of hashes having first been tokenized. An undirected network consisting of two categories of node, *users* and *videos*, is then generated. An edge is created between a user and a video if at least one comment has been posted by the user to the video, where the edge weight represents the number of comments in question. To capture the relationship between the users involved in a particular spam campaign, undirected and unweighted edges are created between user nodes based on the similarity of their associated comment hash sets. Users having only a single video-node neighbor are filtered at this point. This network has a *vocabulary* of 162 motifs up to five nodes—see examples in Figure 6. The set of 162 motifs comprises all possible motifs of between three and five nodes, where edges can exist between users and videos, users and users but not between videos.

Finally, an approximate labeling of the user nodes is performed, where users are labeled as spam users if at least one of their comments is flagged as spam through the YouTube *spam hint* property. As the name suggests that the *spam hint* signal is very noisy as it is often used maliciously against users who are not spammers while, on the other hand, some real spam comments are not flagged. In our annotation of the data, all remaining users are labeled as regular users. Although this can lead

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

to label inaccuracies, the results shown later in this paper demonstrate that such inaccuracies will be perceivable.

5.2 Identifying typical spam accounts

We performed an initial experiment that tracked two spam campaigns which we had discovered following manual analysis of the data covering activity between November 14, 2011 and November 17, 2011. Two distinct campaign strategies were in use, i.e. a small number of accounts each commenting on many videos (Campaign 1), and a larger number of accounts each commenting on few videos (Campaign 2). In order to track the campaign activity over time, a 72-hour period was split into 12 windows of six hours each. For each of these windows, a network of user and video nodes was derived using the process described in the previous section. Network motif profiles were generated for each ego (user), to which principal component analysis (PCA) was then applied in order to produce two-dimensional spatializations of the user nodes, using the first two components. These spatializations acted as the starting point for the analysis of activity within the 12 time windows.

These spatializations based on the full motif set are similar to that shown on the right in Figure 7. Spammers were identified by examining outliers in the PCA spatializations, and confirming these users as belonging to the spam campaigns by analyzing their posted comments. A third spam campaign (Campaign 3) was also active in this 72-hour period. Its strategy appeared to use a large number of users to each post almost identical messages to a small ($\sim$1) number of videos.

This manual process (aided by the PCA) produced labeled examples of three types of spam users. Having generated these three labeled training sets, we ranked the motifs using information gain to determine those that discriminated between spammers and regular users. A selection of 21 motifs was made from those that generated the highest information gain values, seven for each of the three campaigns. These motifs are shown in Figure 6.

5.3 Evaluation

Having selected the 21 discriminating motifs using the November data, a series of experiments was run to analyze the spam activity in the period from December 29, 2011 to January 1, 2012, as it was known that all three campaigns were active during this time. For each six-hour window, motif ratio profiles were generated from the corresponding video and user networks. As the PCA spatializations are used as the starting point for our analysis, we demonstrate the effectiveness of the discriminating motifs in the detection of spam campaigns with the comparison of two spatializations for Window 5 (12 am to 6 am, December 30, 2011); one created using the motif profiles generated from the enumeration of the selected 21 motifs, and the other with profiles generated from all motifs within the network. These spatializations can be seen in Figure 7.

It is clear that the motif-based characterization of the ego-networks seems to be effective in revealing the users involved in the spam campaigns. It is worth remembering that the vast majority of users are regular and appear as overlapping

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

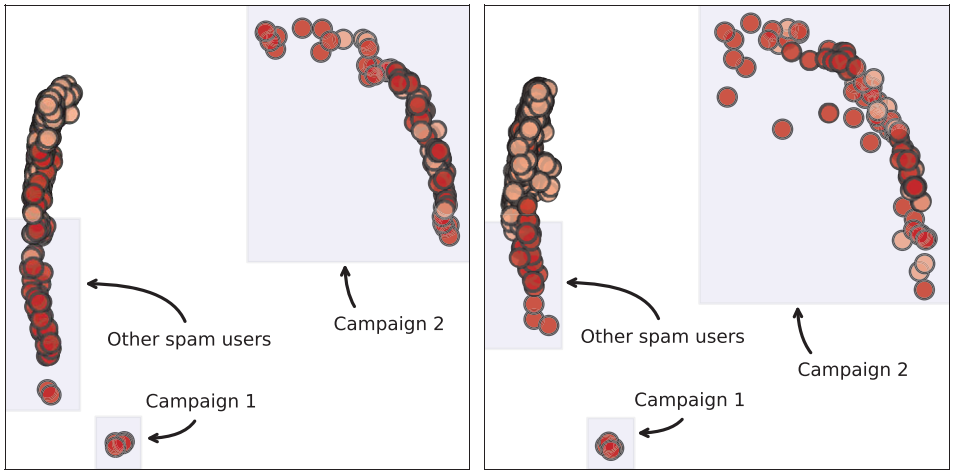


Fig. 7. Spatialization of the first two principal components of the normalized ratio profiles for the YouTube comment networks, Window 5 (12 am to 6 am, December 30, 2011) with 21 motifs (left) and all (161) motifs (right). Dark nodes are users with comments marked as spam, lighter nodes are all other users. Most users are regular, and appear as overlapping points at the top of the leftmost cluster. Both spam campaigns are highlighted. (color online)

points at the top of the leftmost cluster. The spammers are then spread out across the spatialization.

Apart from the highlighted campaign clusters, other spam nodes in the spatializations have been correctly marked as such. For example, those users toward the bottom-left appear to be mostly individual spam accounts having similar behavior to the strategy of Campaign 1, but on a smaller scale. They include users encouraging channel views, i.e. *promoters* (Benevenuto et al., 2009), and also a number of users belonging to a separate campaign. Given the similarity of the spatializations, we can assume that the use of a subset of discriminating motifs in campaign detection can be just as effective as the results obtained when using all motifs. For this application, it seems that a vocabulary of some tens of motifs is required to characterize the ego-networks adequately.

6 The Prosper marketplace

The final example presents publicly available data from Prosper.com, a peer-to-peer lending network.[1] As of August 2012, Prosper.com has more than 1.4 million members and over $375 million in funded loans. Borrowers participate in the Prosper.com marketplace by asking for money in the form of listings. Lenders bid on listings specifying repayment terms including interest rates. If enough lenders fund a listing, the listing becomes a loan. Prosper.com rates prospective borrowers according to their creditworthiness. It also maintains borrower and lender groups, endorsements, past listings, bids, and loans. The social structure of the service is evident from the data: a node represents a borrower or lender and a directed-edge represents a fraction of a listing or bid offered by a lender to a borrower.

[1] More information about Prosper.com can be found on their website at http://www.prosper.com

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

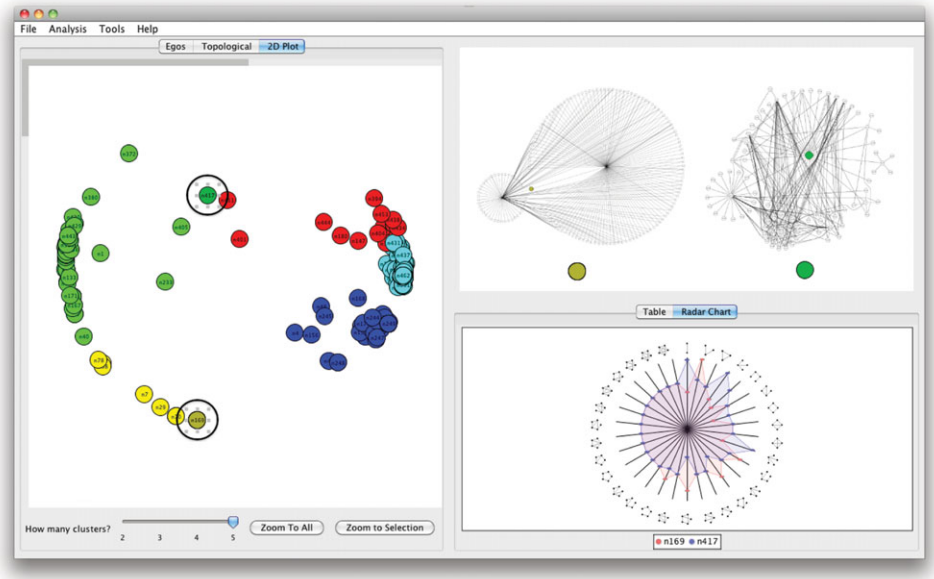


Fig. 8. The activity in the **Prosper marketplace** dataset during April 2010. Two egos are selected in the egocentric spatialization. The selection has been manually annotated with two black circles. Their corresponding egocentric networks and network ratio profiles are shown to the right. The spatialization was produced using principal component analysis. (color online)

In principle, members should be either borrowers or lenders but not both. In fact, about 5% of members act as both borrowers and lenders (Redmond et al., 2012). This behavior is of interest because it may indicate undesirable behavior such as money laundering or arbitrage. This Prosper case study is fundamentally different from the Wikipedia and YouTube examples in the sense that it is inherently unsupervised. In the earlier examples, node labels were available to drive the analysis. By contrast, there is no annotation available to assist with the Prosper analysis so the objective of the assessment is to see if the motif-based characterization produces a spatialization that organizes the nodes in an insightful way.

Figure 8 shows the activity in the Prosper marketplace during April 2010. This figure shows the EgoNav interface (Harrigan et al., 2012), a visualization tool for exploring spatializations of networks based on motif-based characterizations. EgoNav is based on multiple coordinated views (see Figure 8). Each of the three views in the system illustrates a specific aspect of the selected ego(s) and they are coordinated. A selection in any view is highlighted in all other views. The central view of the interface is an egocentric spatialization and is located to the left. This spatialization is computed through network motif analysis PCA. The system interprets the motif profiles as points in a high-dimensional space before projecting them onto a two-dimensional spatialization.

In Prosper in this period, 462 borrowers and lenders agreed upon new loans which were divided into 1,246 fractions. Four hundred and fifty-three of the borrowers and lenders are in a single connected component. The first and second principal components of the network motif profiles (the x- and y-axes of the spatialization)

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

account for 54% and 16% of the variability in the original dataset (see the bar indicators).

Through using EgoNav to explore these networks, it is clear that egos to the left of the spatialization had more nodes and edges than those of the egos to the right. However, the differences between the egos along the *y*-axis of the spatialization are perhaps more interesting. Two representative egos are selected in Figure 8. The radar chart of their network ratio profiles reveals that the ego to the top, when compared with the ego to the bottom, has an egocentric network with relatively fewer low-order network motifs (network motifs with two and three vertices) but relatively more high-order network motifs (network motifs with four and five vertices). This is corroborated by examination of the actual networks in the top-right of the figure. The ego to the bottom of the spatialization has just two neighbors, both of whom are connected to many others. The two neighbors are represented by the nodes at the center of the two circles. However, the ego to the top of the spatialization (shown on the right in the top-right pane) has many more neighbors. These are represented by the nodes in the circle surrounding the ego.

This motif-based analysis suggests that the Prosper member at the bottom of spatialization has an isolated involvement with Prosper while the member at the top is more imbedded in the Prosper community. We note that the differences between, say, the top and bottom egos in Figure 8 can be computed more easily and directly using a combination of node degrees and clustering coefficients. However, the importance and flexibility of the above approach lies in the fact that we did not specify as input the nature of the distinguishing feature(s). Through the exploration of the egocentric spatialization and the network motif profiles we are able to deduce the distinguishing feature(s) and better understand the dataset.

7 Summary and conclusions

The motivation here is to develop a generic mechanism for comparing egocentric networks, without requiring insight into the network features that will be useful for comparison. The idea is to generate a *vocabulary* of all network motifs up to a certain size (say five nodes) and then characterize the network around the ego node using a vector of counts of these motifs. We see from the three example case studies in Wikipedia edit networks, YouTube spam comments, and the Prosper peer-to-peer lending that this strategy is reasonably effective in three quite different application areas.

The first thing to be careful of is that the characterizations used with this strategy do not in fact reduce to something trivial such as an edge count or a measure of network density. The analysis presented in Figure 3 suggests that, in the Wikipedia example, this is not the case. The more complex motifs of size four and five do seem to provide some classification power beyond what is possible with edge, elbow, and triangle counts. Similarly, the example in Figure 6 shows that in the YouTube spam scenario some quite complex motifs indicating different types of collaboration structure have discrimination power.

Perhaps the main shortcoming of this strategy is the fact that the policy of counting all motifs up to a certain size results in a lot of redundant counting. The results in Figure 4 and Table 7 demonstrate this issue. This suggests that in

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

situations where performance is an issue the use of network motif profiling would entail an initial feature selection phase to identify a subset of features that would meet the requirements of the specific task.

The motif-based characterization strategy discussed in this paper is in the spirit of the research originating with Moreno (1934) and developed by Holland and Leinhardt (1976) and Milo et al. (2002). There is some recent work on motif-based characterizations of collaborations in social media that focuses on the temporal characteristics of interactions. Jurgens and Lu (2012) have used temporal sequence motifs to model the dynamics of editor interactions in Wikipedia and Keegan et al.(2012) have introduced the idea of trajectory networks to model the structure and dynamics of collaborations in Wikipedia. It seems clear that the temporal dynamics of interactions contains important information so our next research challenge is to extend our work to capture the dynamics of interactions.

Acknowledgments

This work is supported by Science Foundation Ireland grant no. 08/SRC/I140 (Clique: Graph & Network Analysis Cluster).

References

Allan, Jr., Edward, G., Turkett, Jr., William, H., & Fulp, E. W. (2009). Using network motifs to identify application protocols. *Proceedings of the 28th IEEE Conference on Global Telecommunications (GLOBECOM'09)*, Piscataway, NJ: IEEE Press, pp. 4266–4272.

Anderson, C. J., Wasserman, S., & Crouch, B. (1999). A p* primer: Logit models for social networks. *Social networks*, **21**(1), 37–66.

Antiqueira, L., & da Fontoura Costa, L. (2009). Characterization of subgraph relationships and distribution in complex networks. *New Journal of Physics*, **11**(013058).

Artzy-Randrup, Y., Fleishman, S. J., Ben-Tal, N., & Stone, L. (2004). Comment on "Network motifs: Simple building blocks of complex networks" and "superfamilies of evolved and designed networks". *Science*, **305**(5687), 1107.

Becchetti, L., Boldi, P., Castillo, C., & Gionis, A. (2008). Efficient semi-streaming algorithms for local triangle counting in massive graphs. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*, New York: ACM, pp. 16–24.

Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., & Gonçalves, M. (2009). Detecting spammers and content promoters in online video social networks. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*, New York: ACM, pp. 620–627.

Borgatti, S., & Everett, M. (1989). The class of all regular equivalences: Algebraic structure and computation. *Social Networks*, **11**, 65–88.

Borgatti, S. P., & Everett, M. G. (1992). Notions of position in social network analysis. *Sociological Methodology*, **22**(1), 1–35.

Borgwardt, K. M., & Kriegel, H. P. (2005). Shortest-path kernels on graphs. *Fifth IEEE International Conference on Data Mining*, New York: IEEE, pp. 74–81.

Boykin, P. O., & Roychowdhury, V. P. (2005). Leveraging social networks to fight spam. *Computer*, **38**(4), 61–68.

Brandes, U., Lerner, J., Lubbers, M., McCarty, C., & Molina, J. (2008). Visual statistics for collections of clustered graphs. *Proceedings of the IEEE VGTC Pacific Visualization Symp. (PacificVis'08)*, New York, pp. 47–54.

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In M. Cord & P. Cunningham (Eds.), *Machine learning techniques for multimedia* (pp. 21–49). Berlin: Springer.

Davis, J. A. (1963). Structural balance, mechanical solidarity, and interpersonal relations. *American Journal of Sociology*, **68**, 444–462.

Faust, K. (2007). Very local structure in social networks. *Sociological Methodology*, **37**(1), 209–256.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**(2), 179–188.

Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., & Zhao, B. Y. (2010). Detecting and characterizing social spam campaigns. *Proceedings of the 10th Annual Conference on Internet Measurement (IMC '10)*. New York: ACM, pp. 35–47.

Gärtner, T., Flach, P., & Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives. *Learning Theory and Kernel Machines*, **2777**, 129–143.

Harrigan, M., Archambault, D., Cunningham, P., & Hurley, N. (2012). EgoNav: Exploring networks through egocentric spatializations. *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI 2012)*. New York: ACM, pp. 563–570.

Holland, P. W., & Leinhardt, S. (1976). Local structure in social networks. *Sociological Methodology*, **7**(1).

Horváth, T., Gärtner, T., & Wrobel, S. (2004). Cyclic pattern kernels for predictive graph mining. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, pp. 158–167.

Jurgens, D., & Lu, T.-C. (2012). Temporal motifs reveal the dynamics of editor interactions in Wikipedia. In J. G. Breslin, N. B. Ellison, J. G. Shanahan, & Z. Tufekci (Eds.), *Proceedings of the Sixth International Conference on Weblogs and Social Media (ICWSM 2012)*. Palo Alto, CA: AAAI Press.

Juszczyszyn, K., Kazienko, P., & Musiał, K. (2008). Local topology of social network based on motif analysis. In I. Lovrek, R. Howlett, & L. Jain (Eds.), *Knowledge-based intelligent information and engineering Systems* (pp. 97–105). Lecture Notes in Computer Science, vol. 5178. Berlin/Heidelberg: Springer.

Kalish, Y., & Robins, G. (2006). Psychological predispositions and network structure: The relationship between individual predispositions, structural holes and network closure. *Social networks*, **28**(1), 56–84.

Kamaliha, E., Riahi, F., Qazvinian, V., & Adibi, J. (2008). Characterizing network motifs to identify spam comments. *Proceedings of the 2008 IEEE International Conference on Data Mining Workshops*, Washington, DC: IEEE Computer Society, pp. 919–928.

Kashima, H., Tsuda, K., & Inokuchi, A. (2004). Kernels for graphs. *Kernel Methods in Computational Biology*, **39**(1), 101–113.

Keegan, B., Gergle, D., & Contractor, N. (2012). Staying in the loop: Structure and dynamics of Wikipedia's breaking news collaborations. *Proceedings of the 8th International Symposium on Wikis and Open Collaboration,* Linz, Austria.

Krause, J., Croft, D. P., & James, R. (2007). Social network theory in the behavioural sciences: Potential applications. *Behavioral Ecology and Sociobiology*, **62**(1), 15–27.

Lubbers, M., Molina, J., Lerner, J., Brandes, U., Ávila, J., & McCarty, C. (2010). Longitudinal analysis of personal networks: The case of Argentinean migrants in Spain. *Social Networks*, **32**(1), 91–104.

Luo, B., Wilson, R. C., & Hancock, E. R. (2003). Spectral embedding of graphs. *Pattern Recognition*, **36**(10), 2213–2230.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science*, **298**(5594), 824–827.

https://doi.org/10.1017/nws.2013.12 Published online by Cambridge University Press

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., & Alon, U. (2004). Superfamilies of evolved and designed networks. *Science*, **303**(5663), 1538.

Moreno, J. L. (1934). *Who shall survive? A new approach to the problem of human interrelations.* New York: Nervous and Mental Disease Publishing Co.

O'Callaghan, D., Harrigan, M., Carthy, J., & Cunningham, P. (2012). Network analysis of recurring YouTube spam campaigns. In J. G. Breslin, N. B. Ellison, J. G. Shanahan, & Z. Tufekci (Eds.), *Proceedings of the Sixth International Conference on Weblogs and Social Media (ICWSM 2012)*. Palo Alto, CA: The AAAI Press.

Paton, K. (1969). An algorithm for finding a fundamental set of cycles of a graph. *Communications of the ACM*, **12**(9), 514–518.

Pržulj, N. (2007). Biological network comparison using graphlet degree distribution. *Bioinformatics*, **23**(2), e177–e183.

Ramon, J., & Gärtner, T. (2003). Expressivity versus efficiency of graph kernels. *First International Workshop on Mining Graphs, Trees and Sequences*, Osaka, Japan, pp. 65–74.

Redmond, U., Harrigan, M., & Cunningham, P. (2012). Mining dense structures to uncover anomalous behaviour in financial network data. In M. Atzmueller, A. Chin, D. Helic, & A. Hotho (Eds.), *Modeling and mining ubiquitous social media* (pp. 60–76). Lecture Notes in Computer Science, vol. 7472. Berlin, Heidelberg: Springer.

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007a). An introduction to exponential random graph (p*) models for social networks. *Social Networks*, **29**(2), 173–191.

Robins, G., Snijders, T., Wang, P., Handcock, M., & Pattison, P. (2007b). Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, **29**(2), 192–215.

Saul, Z. M., & Filkov, V. (2007). Exploring biological network structure using exponential random graph models. *Bioinformatics*, **23**(19), 2604.

Shervashidze, N., Vishwanathan, S. V. N., Petri, T., Mehlhorn, K., & Borgwardt, K. (2009). Efficient graphlet kernels for large graph comparison. *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS'09)*, Cambridge, MA.

Stoica, A., & Prieur, C. (2009). Structure of neighborhoods in a large social network. *Proceedings of the International Conference on Computational Science & Engineering (CSE'09)*, New York, pp. 26–33.

Vazquez, A., Dobrin, R., Sergi, D., Eckmann, J. P., Oltvai, Z. N., & Barabási, A. L. (2004). The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(52), 17940.

Wellman, B. (1993). An egocentric network tale: Comment on Bien *et al. Social Networks*, **15**, 423–436.

Welser, H., Gleave, E., Fisher, D., & Smith, M. (2007). Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure*, **8**.

Wernicke, S., & Rasche, F. (2006). FANMOD: A tool for fast network motif detection. *Bioinformatics*, **22**(9), 1152.

White, H., Boorman, S., & Breiger, R. (1976). Social structure from multiple networks – blockmodels of roles and positions. *American Journal of Sociology*, **81**, 730–780.

Xie, Y., Yu, F., Achan, K., Panigrahy, R., Hulten, G., & Osipkov, I. (2008). Spamming botnets: Signatures and characteristics. *Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication (SIGCOMM '08)*. New York: ACM, pp. 171–182.