EgoNav: Exploring Networks through Egocentric Spatializations

Martin Harrigan, Daniel Archambault, Pádraig Cunningham, and Neil Hurley
School of Computer Science and Informatics
University College Dublin, Ireland
{martin.harrigan,daniel.archambault,padraig.cunningham,neil.hurley}@ucd.ie

ABSTRACT

EgoNav is a visual analytics system that characterizes egos based on the relationship structure of their egocentric networks and presents the results as a spatialization. An ego, or individual node in a network, is most closely related to its neighbors, and to a lesser degree, to its neighbor's neighbors. For example, in social networks, people are closely related to their friends and family. In financial networks, the affairs of borrowers and lenders are more closely tied to each other. In fact, the relationship structure surrounding an ego, or an egocentric network, can provide characteristic information about the ego itself. Using network motif analysis and dimensionality reduction techniques, the system places egos in similar areas of a spatialization if their egocentric networks are structurally similar. This view of a network discriminates between the various classes of typical and exceptional egos. We demonstrate its effectiveness using appropriate synthetic datasets, real-world mobile phone call and peer-to-peer lending datasets. We subsequently elicit user feedback from experts involved in the investigation of financial fraud to assess the tool's applicability in this domain.

Categories and Subject Descriptors

H.5.m [**Information Interfaces and Presentation**]: Miscellaneous; G.2.2 [**Discrete Mathematics**]: Graph Theory—*Graph algorithms*

General Terms

network motif analysis, spatializations, visual analytics

1. INTRODUCTION

An egocentric network comprises an ego, or individual node in a network, and nodes that are closely related to the ego along with all edges between those nodes. An egocentric network describes the local view of an ego in a network. To analyze the egos in a network, it is necessary to extract, analyze and visually summarize their egocentric networks. However, with a few notable exceptions [5, 28, 18], current state-of-the-art network analysis and visualization tools focus on analyzing and visualizing either the entire network or individual egocentric networks but fail to visually summarize a collection of egocentric networks.

An egocentric network provides characteristic information about the ego itself. For example, consider a bank account in a financial transaction network. If the account is typical, its egocentric network should be a member of a class of typical accounts to which it is structurally similar. On the other hand, consider a bank account involved in *smurfing*, the splitting of large transactions into multiple smaller transactions, each of which is below a limit above which financial institutions must report. This problem was posed to us by an industrial partner in the financial services compliance industry. Assuming the incidence of smurfing is relatively low, the bank account's egocentric network should be exceptional.

EgoNav is a system for analyzing and visually summarizing a collection of egocentric networks. We characterize the structure of egocentric networks using *network motif analysis* [25]. Network motif analysis exhaustively counts the number of network motifs up to a certain size in a network. For large networks, this can be prohibitive, but for a collection of egocentric networks we can divide and parallelize the computation. We do not specify as input a typical or exceptional ego. The only inputs to EgoNav are the entire network, the longest shortest-path distance k from an ego to every node in the ego's egocentric network (the radius), and the network motifs we would like to count. The counts for each egocentric network are adjusted for scale to produce *network ratio profiles*. The network ratio profiles can be interpreted as points in a high-dimensional space. They are projected onto a 2-dimensional spatialization using *principal component analysis* (PCA) or *locality preserving projections* (LLE [26] and Isomap [29]). The spatialization encodes the similarities and differences amongst the egocentric networks. Furthermore, clusters of egos represent broad classes of egos with structurally similar egocentric networks.

Our main contribution is a visual analytics system that combines an egocentric variation of network motif analysis and dimensionality reduction to produce an aggregated view of a network. The system, EgoNav, allows a user to visually inspect egocentric networks, network ratio profiles, and a spatialization of the egos based on the structure of their egocentric networks. The various views are coordinated allowing a user to select an ego in one view and examine its properties in another. A user can also compare, for example, network ratio profiles through multiple selections to help identify the distinguishing features of a collection of egocentric networks. We trialed the system with experts in the financial services compliance industry with a view to detecting fraud in network datasets.

2. RELATED WORK

In this section, we review structural similarity (Sect. 2.1), network motif analysis (Sect. 2.2), and techniques for analyzing and visualizing collections of egocentric networks (Sect. 2.3).

AVI '12 May 21-25, 2012, Capri Island, Italy

2.1 Structural Similarity

The structural similarity of two nodes can be measured in two ways: two nodes are *structurally equivalent* [19] if they share many of the same neighbors; two nodes are *regularly equivalent* if they are connected to other nodes that are themselves structurally similar. Leicht et al. [17] formulate a measure of the latter. They compute a weighted count of the number of paths of all lengths between the nodes in question. This generalizes the notion of structural equivalence in that paths of all lengths are considered. They experimented with their similarity measure using random stratified networks, Roget's Thesaurus, and a friendship network of high school students and found that their measure produced favorable results when compared with other similarity measures [3].

Structural similarity can also be used to classify entire networks. Given a set of random (planted partition [23]) network models and a set of random networks generated using those models, Brandes et al. [6] provide an effective heuristic for partitioning the set of random networks such that two networks are in the same part if and only if they are generated using the same random network model. Their heuristic relies on the fact that the adjacency matrices of two networks generated using the same random planted partition network model have (with high probability) similar spectra.

The above methods are applied at the node and network level respectively. However, we require a structural similarity measure at the egocentric level. For this purpose, we use network motif analysis (see the following subsection and Sect. 3 for more details).

2.2 Network Motif Analysis

Network motif analysis represents the structural or topological properties of a network using network motif counts [24]. It essentially counts all occurrences of each network motif. Milo et al. [24] use *network ratio profiles* to compare networks of varying sizes. Using a correlation coefficient matrix, they identify several families of networks that have similar network ratio profiles, such as biological information-processing networks, WWW networks, social networks and autonomous systems networks.

Koschützki et al. [16] formulate a number of network motif-based centrality measures. They rank the nodes of the E. Coli transcriptional network using each centrality measure. They claim that network motif-based centrality measures identify genes that are import regulators which are overlooked by local (*e.g.* out-degree) and global (*e.g.* betweenness) centrality measures.

Network motif analysis is generally concerned with entire networks and global counts. However, we adapt this approach to egocentric networks by computing local counts – a network motif must be present in an egocentric network and incident with the ego before it is counted (see Sect. 3 for more details). The results of Koschützki et al. [16] lend support to this approach by using local counts of network motifs to assess the importance of nodes.

Several visualization systems use network motif analysis, such as `MAVisto` [27, 15], `FANMOD` [35] `SNAVI` [21] and Huang et al. [12]. However, these are concerned with network motif analysis on entire networks and have no egocentric capabilities.

2.3 Egocentric Analysis and Visualization

Egocentric analysis has a long-history within the field of social network analysis [36, 4, 33]. For example, a recent work [20] 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 were visualized using clustered networks [5] where the size of four nodes encode the number of people in each of four groups (origin, fellows, host and transnationals) and the thickness of the edges quantify the amount of communication between the groups.

Welser et al. [34] 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 poster: answer people, discussion people, and disruptors. Similarly, Stoica and Prieur [28] 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 [1] present a methodology for analyzing non-overlapping subnetworks, their interrelationships, and their distribution in a network. They analyze four random network models 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 [34, 28, 1] 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.

Although not directly related to egocentric visualization, there are three other systems we would like to mention. Von Landesberger et al. [31] use a *self-organizing map* (SOM) to cluster networks into a grid of prototypical networks. They compute a variety of topological features for the networks. The user weights the features appropriately and the system produces a SOM layout. Each cell represents a subset of similar networks. Bezerianos et al. [2] describe `GraphDice`, system for visualizing the attribute values of the nodes in a network. A scatterplot matrix provides an overview of all possible combinations of attributes and a 2-dimensional projection provides a detailed view. Freire et al. [9] use a tabular interface to inspect many networks simultaneously. Their tool, `ManyNets`, also provides summary statistics for a selected subset of the networks. Our interface is also inspired by `iPCA` [13], an interactive interface that opens up the black box of PCA.

3. VISUAL ANALYTICS SYSTEM

The motivating task behind `EgoNav` was posed to us by an industrial partner in the financial services compliance industry. In a typical use case, they require a method of identifying instances of suspicious patterns that potentially signal internal fraud perpetrated by an employee. They describe one suspicious pattern as small densely connected subnetworks in a financial transaction network but they note that other suspicious patterns may also exist. In fact, they require a method of identifying suspicious patterns that they are not already familiar with. It is a relatively simple task to implement a method of identifying small densely connected subnetworks but it is a much more difficult task to identify suspicious patterns that are, as yet, unknown. We believe that the visual analytics system below tackles this task. Briefly, given a network, we extract the egocentric networks, enumerate the network motifs, compute the network motif profiles and the network ratio profiles, and project the network ratio profiles, represented as points in a high-dimesional space, onto a 2-dimensional spatialization. This approach places egos with similar egocentric networks in similar areas of the spatialization.

3.1 The EgoNav Interface

The design and implementation of EgoNav is based on multiple coordinated views (see Fig. 1). Each of the three views in the system illustrate a specific aspect of the selected ego(s) and are coordinated. A selection in any view is highligted 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 and dimensionality reduction, as described in Sect. 3.2 and 3.3. The system interprets the network ratio profiles as points in a high-dimensional space before projecting them onto a 2-dimensional spatialization. The view allows for panning and zooming. Two bar indicators are situated in the top left corner of the egocentric spatialization. The length of the bars indicate the percentage of variability captured by the spatialization.

Egos are colored using a k-means clustering. The value of k can be set by the user using a slider located at the bottom of the interface. The correct choice of k is often ambiguous; it depends on the shape and scale of the dataset and the desired clustering resolution of the user. The clustering is computed on the projected data in the 2-dimensional spatialization. The correlations in the high-dimensional space are removed using dimensionality reduction techniques. The coloring serves to re-enforce the clustering rather than introduce any new information. We chose this encoding for our central view to help users estimate the number of egos in each cluster and due to performance results relating to the counting of colored points on a 2-dimensional spatialization [30].

When a point or group of points representing egos are selected in the egocentric spatialization, their corresponding details appear in the top right and lower right portions of the interface. An egocentric view, located at the top right, displays the topology of the egocentric network(s) of the selected ego(s). The network ratio profile view, located at the lower right, displays either a table or a radar chart of the network motif profile(s) of the selected ego(s). The values in the table range from -1 to 1: 0 indicates that a particular network motif occurs as often as you would expect given its prevalence in the network; a negative or positive value indicates that a particular network motif is under- or over-represented respectively. The rows of the table can be sorted by increasing or decreasing value to show the most under- or over-represented network motifs. The radar chart represents each network motif as a spoke. Each shaded area represents a *network ratio profile* (see Sect. 3.3). It intersects the spokes at a distance from their midpoints that is proportional to the value for the corresponding network motif. This allows a user to compare multiple network ratio profiles.

3.2 Network Motif Enumeration

The *egocentric network*, or *k-neighborhood subnetwork*, of an ego u is the subnetwork induced by the set of nodes that have shortest-path distance at most k from u. In this paper we set $k = 2$. We consider the egocentric network of each node in turn.

We construct an ordered list of all connected networks with at most l nodes up to isomorphism using `geng` from the `nauty` package [22]. In this paper we set $l = 5$, and therefore the ordered list has 30 networks or network motifs. For each egocentric network we compute a *network motif profile*: a 30-element vector where each entry is the number of instances of the corresponding network motif in the ordered list that are incident with the ego. We compute these counts using `GraphGrepSX` [11], a tool that solves the subgraph isomorphism problem using enumerated paths as index features. This is a time-consuming process. However, for large datasets we can process the egocentric networks in parallel and/or reduce both k and l.

For directed networks, we need to generate all connected directed networks with at most l nodes up to isomorphism. Similarly for colored networks, *i.e.* networks where the node and edge attribute values have a small number of discrete categorizations, we need to generate all connected colored networks with at most l nodes up to isomorphism. These requirements greatly increase the number of network motifs making the network motif profiles difficult to compute. We are investigating online algorithms for maintaining network motif profiles for directed and colored networks over time. However, the system in this paper handles multiedge, undirected networks only.

3.3 Scaling and Correlation Adjustments

For each network motif profile, we compute a *network ratio profile* [24]: a 30-element vector where each entry is the *normalized ratio* of the corresponding entry in the network motif profile. The ratio profile rp of an egocentric network is computed using

$$rp_i = \frac{nmp_i - \overline{nmp_i}}{nmp_i + \overline{nmp_i} + \varepsilon}$$

where nmp_i is the i^{th} entry of the network motif profile, $\overline{nmp_i}$ is the average of the i^{th} entry of all of the network motif profiles, and ε is a small integer that ensures that the ratio is not misleadingly large when the network motif appears very few times in all of the egocentric networks. To adjust for scaling the normalized ratio profile nrp of an egocentric network is computed using

$$nrp_i = \frac{rp_i}{\sqrt{\sum rp_j^2}}.$$

A normalized ratio measures the abundance of a network motif in each individual egocentric network relative to all the egocentric networks; it is similar to a z-score. It does not require the construction of a random network ensemble as found in related approaches [24].

There are correlations between the elements of a network ratio profile [10]. To adjust for these, we use a dimensionality reduction technique such as principal component analysis (PCA), locally linear embedding (LLE) [26] or Isomap [29]. In the case of PCA, we use two bar indicators in the top left corner of the egocentric spatialization to indicate the magnitudes of the eigenvalues for the corresponding axes. When the percentage of variability accounted for by either axis is small, indicated by a short bar indicator, the positioning of the points along that axis is not significant. In these cases it may be preferable to use one of the two locality preserving techniques.

4. CASE STUDIES AND FEEDBACK

To support our methodology, we explored, in collaboration with our industrial partner in the financial services compliance industry, four datasets. We generated two synthetic datasets using generators provided by our industrial partner and we derived two real-world datasets from the MIT Reality Mining project [7] and the Prosper Marketplace. In the subsections below, we describe the significance of each dataset for our industrial partner, any insights into the datasets garnered through the explorations, and feedback regarding the interface and methodology.

We performed the explorations below with two analysts from our industrial partner. Both were experienced in techniques and systems for complying with regulatory requirements and tackling crime in the financial services industry, for example, money laundering, multi-channel fraud, and internal fraud. They described existing solutions as being predominantly rule-based (see, for example, Khan et al. [14]). The set of rules that apply to a particular case guide subsequent human investigation. The rules are often based on experience. However, these methods only detect certain

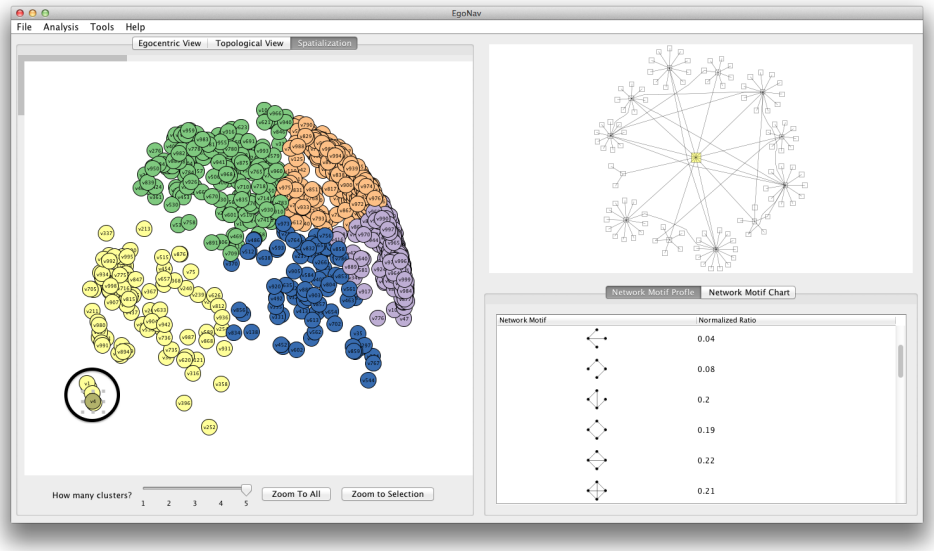


Figure 1: Exploration of a single $1,000$-node network from a synthetic dataset. This dataset was generated using a variation of the Erdös-Rényi model [8] (see Sect. 4.2). On the left is the 2-dimensional spatialization. The five egos belonging to the five node clique have been manually annotated with a black circle. The exceptionality of these egos is apparent from their position. One of the egos is selected. On the top right is the egocentric network for the selected ego. On the lower right is the network ratio profile for the selected ego.

types of fraud; their capabilities are usually limited since the detection relies on simple, built-in rules defined by domain experts. The fraud investigators and auditors require an in-depth understanding of the rules, their syntax and semantics, when tackling new and heretofore unseen types of fraud.

We introduced egocentric network motif analysis to the analysts. They worked in unison when tackling the datasets below. They drove the user interface themselves but were allowed to ask questions to the experimenter at all stages. We also documented any improvements necessary to integrate `EgoNav` into the workflow of a fraud investigator.

4.1 The MIT Reality Mining Project

We first explored the **MIT Reality Mining** dataset [7]. It comprised mobile phone call and SMS records over a 296-day period between 100 unique mobile phones. The dataset was a subset of a much larger dataset comprising communication, proximity, location, and activity information involving 100 subjects at MIT over the course of the 2004-2005 academic year. By letting a node represent a user, or more specifically a mobile phone, and an edge represent a mobile phone call or SMS between two mobile phones, we were able to visualize and explore the data using `EgoNav`. Figure 2 shows an egocentric view and a global view of the network.

The purpose of this dataset was to introduce the interface and methodology of `EgoNav` to our analysts. We explained the concepts of egocentricity, network motif analysis, structurally similar communication patterns and egocentric spatializations together with the elements of interface. We showed that the global view identified two large communities; this is a known artifact of the dataset [7]. We also showed that the egocentric view identified two clusters but that these clusters did not correspond to the two communities in the global view. Instead, they corresponded to the core mobile phone users and the peripheral mobile phone users. The local view clusters egos based on the structural similarity of their corresponding egocentric networks. Therefore all egos that are core mobile phone users, regardless of which community they belong to, cluster together. We further divided the peripheral mobile phone users into an inner periphery (the green egos) and an outer periphery (the blue egos).

4.2 The Synthetic Datasets

A motivating challenge was presented to us by our industrial partner as follows: 'Given a set of transactions detailing a source account, a destination account, a responsible employee and a branch, can we identify instances of suspicious patterns that potentially signal internal fraud perpetrated by an employee? Specifically, one pattern that we consider suspicious involves many transactions between a limited set of accounts performed by a single employee. However, there may be others that are also suspicious.' The question is somewhat open-ended. We easily identified instances of the given suspicious pattern by searching for cliques whose nodes and edges had the appropriate attributes but it was difficult to extend this approach to handle suspicious patterns that were not known *a priori*. We hypothesized that the methodology followed by `EgoNav` identifies instances of the given suspicious pattern but is also flexible enough to identify instances of other suspicious patterns. To illustrate this we generated two synthetic datasets based on two random network models and inserted a small clique in each.

The **ER** dataset includes 50 random networks generated using the Erdös-Rényi model [8]. Each network contains $1,000$ nodes and $5,000$ edges. Furthermore, five nodes were chosen at random and augmented with additional edges to create a clique. The **WS** dataset also includes 50 random networks generated using the Watts-Stogatz model [32]. Each network contains $1,000$ nodes and $5,000$ edges. Again, five nodes were chosen at random and augmented with additional edges to create a clique. Our industrial partner uses a similar generator when testing their systems; given a number of accounts, a number of branches, an average number of

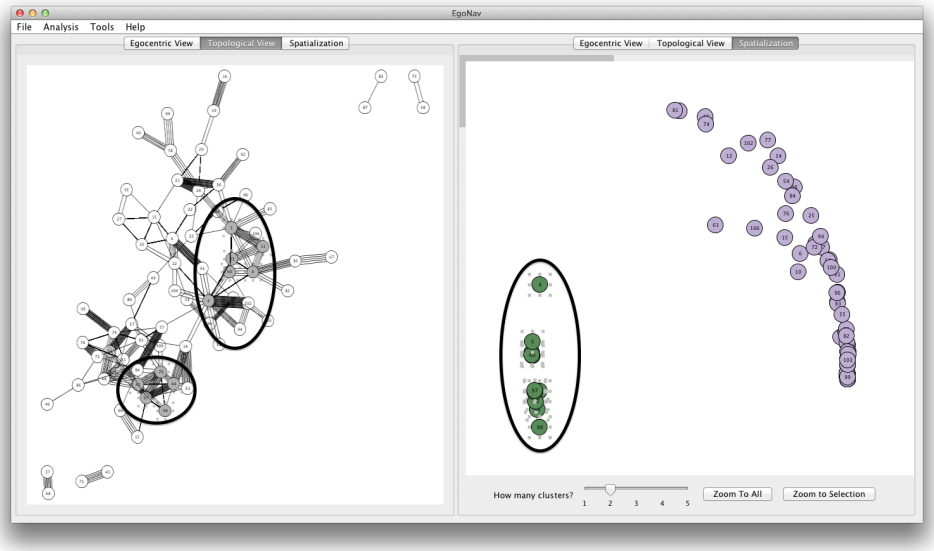


Figure 2: A comparison of an egocentric view and a global view of the MIT Reality Mining dataset. Both views identify two large clusters or communities: the egocentric view (left side) discriminates between core and peripheral mobile phone users whereas the global view (right side) identifies two communities in the more traditional sense. The selected egos have been manually annotated with black circles.

employees per branch, an average number of transactions per day, and an average number of instances of the above suspicious pattern per day, their generator produces a list of fictitious financial transactions. Their generator essentially uses the Erdös-Rényi model along with observed empirical distributions when populating node and edge attributes, for example, the 'type' of an account or the value of a transaction.

The analysts understood the correspondence between the data generated with our generators and their own. They easily identified the cliques in the **ER** dataset using the egocentric spatialization. Figure 1 shows a typical 1,000-node network from the **ER** dataset. We note that the cliques were not easily identifiable in the corresponding global views produced by a force-directed algorithm. Figure 3(a) shows the network ratio profiles of the egocentric networks of the five egos in the clique. All values in the network ratio profiles are relatively high, especially for the higher-order network motifs. This is what makes the corresponding egos exceptional. At this point in our discussion with the analysts, we uncovered a potential source of confusion: one analyst stated that the egos were obviously exceptional when visually inspecting the size and the density of their corresponding egocentric networks. However, it is important to note that the exceptionality of an ego cannot be determined in isolation; it must be compared with all other egos in order to assess whether a particular network motif is under- or over-represented.

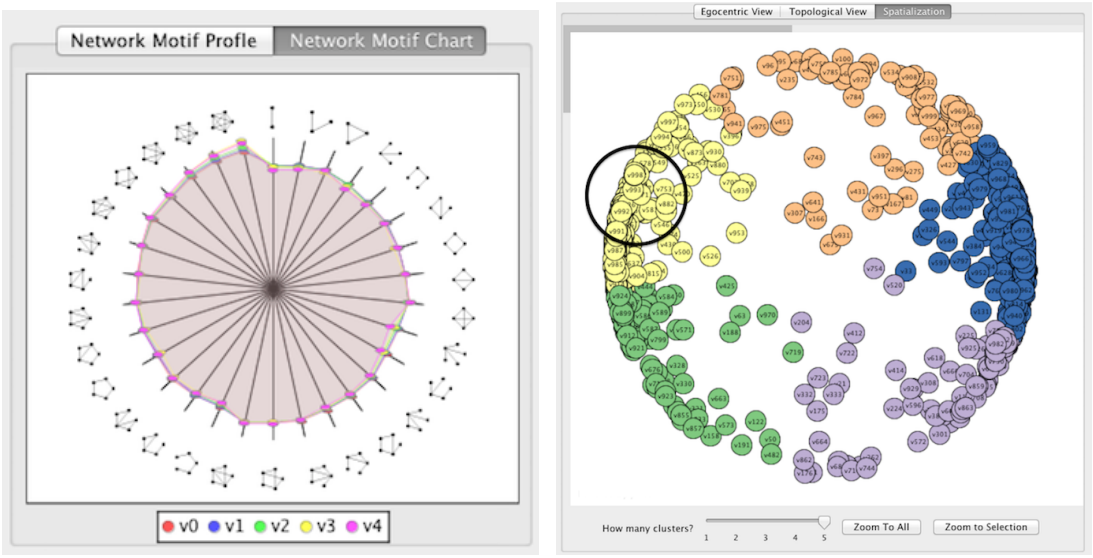
The analysts failed to identify the cliques in the **WS** dataset. Figure 3(b) shows a typical 1,000-node network from the **WS** dataset. The clique was not identifiable through the egocentric spatialization. This was due to the increased clustering coefficient found in networks from the **WS** dataset compared to those from the **ER** dataset. The egos in the clique were no longer considered exceptional. This was true for all 50 networks in the dataset. The networks in both Fig. 1 and Fig. 3(b) have the same number of nodes and edges. However, their differing structure means that an egocentric network considered exceptional in one is typical in another. Both analysts understood the difficulty; they suggested that a combination of filtering techniques based on node and edge attributes and weighting the individual values in the network ratio profiles in order to identify certain types of egocentric networks might help.

4.3 The Prosper Marketplace

Finally, the analysts considered a publicly available dataset from the financial domain. The **Prosper Marketplace**[1] dataset is derived from a peer-to-peer lending or social lending service – borrowers ask for money in the form of listings and 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 an edge represents a fraction. We note that lenders can also be borrowers and *vice versa* and therefore the network is not necessarily bipartite.

The analysts had no prior knowledge of this dataset. We explained the workings of Prosper.com and asked the analysts to explore the dataset for themselves. The analysts considered the dataset in monthly snapshots. Figure 4 shows the activity in the **Prosper Marketplace** dataset during April 2010. 462 borrowers and lenders agreed upon new loans which were divided into 1,246 fractions. 453 of the borrowers and lenders are in a single connected component. The first and second principal components of the network ratio profiles (the x- and y-axes of the spatialization) account for 54% and 16% of the variability in the original dataset (see the bar indicators). Through visual inspection, the analysts observed that the egocentric networks of egos to the left of the spatialization had more nodes and edges than those of the egos to the right. However, the difference between the egos along the y-axis of the spatialization was deemed more interesting. Two representative egos are selected in Fig. 4. The radar chart of their network ratio profiles reveal

[1] http://www.prosper.com



(a) (b)

Figure 3: The analysts used EgoNav to inspect the ER and WS datasets. (a) When they selected all five nodes of the clique for the network in Fig. 1, they were able to view a summarization of the five network ratio profiles using a radar chart. They could see that the five egos had broadly identical network ratio profiles and were therefore structurally similar. (b) A single $1,000$-node network from the WS dataset. The five egos belonging to the five node clique have been manually annotated with a black circle. Due to the increased clustering coefficient found in networks from the WS dataset compared to those from the ER dataset, the egos belonging to the clique are not considered exceptional.

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 the small multiples representation of the two corresponding egocentric networks. The ego to the bottom of the spatialization (to the left of the small multiples representation) 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 (to the right of the small multiples representation) has many more neighbors. These are represented by the nodes in the circle surrounding the ego. The analysts repeatedly selected representative egos from the extreme sides of the spatialization and considered their corresponding network ratio profiles and the 'shape' of their corresponding egocentric networks. When they understood the discriminating feature for an axis they moved to the next axis and repeated the process. They suggested that a third or fourth axis might also be relevant.

We note that the differences between, say, the top and bottom egos in Fig. 4 can be computed more easily and directly using, say, 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 ratio profiles the analysts were able to deduce the distinguishing feature(s) and better understand the dataset.

4.4 Domain Specific Feedback

Much of the feedback provided by the analysts was domain-specific and related to the integration of `EgoNav` into the workflow of a fraud investigator. Firstly, they suggested less regard for technical terminology and more regard for the purpose and business meaning of the components of the interface. For example, they suggested substituting 'account behavior' for network motif profile and 'score' for normalized ratio, making the radar chart the default representation for a network ratio profile and using icons to represent node attributes like account type. They also suggested exaggerating the distances between where the shaded areas intersect the spokes of the radar chart and their corresponding midpoints using a logarithmic scale. They differentiated between two types of user: a routine user who uses `EgoNav` according to a manual or standard operating procedure and a power user who performs exploratory analyses, identifies and categorizes new types of fraud and maintains a library of representative instances of fraud that a routine user can later consult. However, both need a shared vocabulary.

Secondly, they discussed integration with existing rule-based systems. The easiest point of integration is to allow flags raised by a rule-based system to be visible from within `EgoNav` as attributes of a node or edge. They differentiated between two types of analyses: offline analyses of logged financial transactions and online, real-time analyses. Many rule-based systems provide for both. However, `EgoNav` caters for offline analyses only. The network motif profiles, and hence, the network ratio profiles and egocentric spatializations, are computed offline. However, we are investigating online algorithms for maintaining network motif profiles and interface components for displaying changing network ratio profiles and egocentric spatializations.

Thirdly, the analysts suggested that all parameters, *e.g.* the size of the egocentric networks and the size of the largest network motif (k and l in Sect. 3.2 respectively) should be configurable. They understood the trade-off between increasing the values of the parameters and the computational overhead required to consider larger egocentric networks and larger network motifs.

`EgoNav` ignores all node and edge attributes and directional-

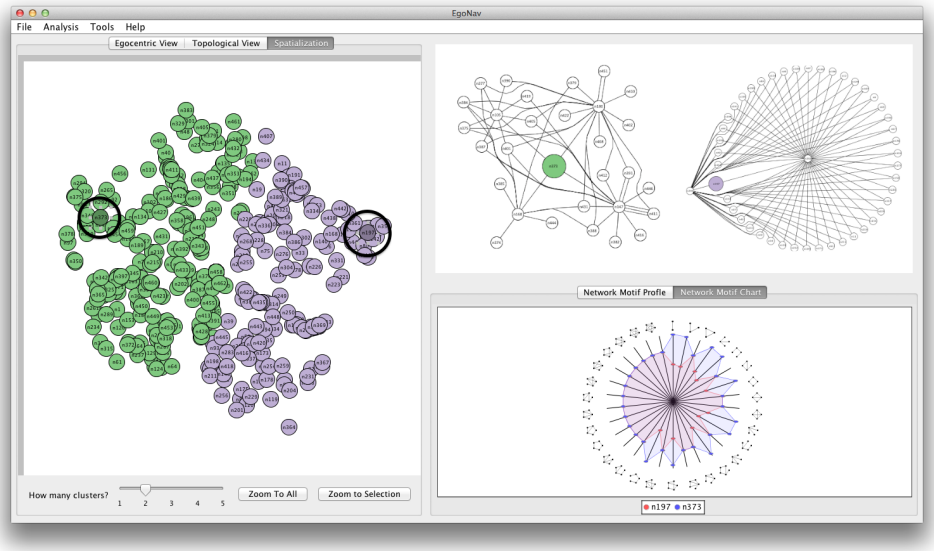


Figure 4: The activity in the Prosper Marketplace dataset during April 2010. Two egos are selected in the egocentric spatialization. The selection has been manually annotated with two black circles. Their corresponding egocentric networks and network ratio profiles are shown to the right. The spatialization was produced using Isomap.

ity. However, these features can be handled by extending the list of network motifs. Temporal information can be also handled by maintaining network motif profiles over time. Both the **MIT Reality Mining** and the **Prosper Marketplace** datasets have node and edge attributes, are directed and have temporal information. An extended list of network motifs would enable `EgoNav` to discover a wider range of discriminating features.

5. CONCLUSIONS AND FUTURE WORK

We have presented `EgoNav`, a system for exploring networks through egocentric spatializations. Using a combination of network motif analysis at the egocentric level and dimensionality reduction using PCA, LLE and Isomap, the system produces spatializations that reflect the similarities and differences amongst all egocentric networks in a network. The interface has multiple coordinated views to better represent the various aspects of the egocentric networks. The views allow a user to select an ego in one view and examine its properties in another. For example, the system includes a view of the topology of the selected egocentric networks and a view for comparing their network ratio profiles. The egocentric spatialization is the central view in the system. It includes bar indicators to show the significance of the two axes and a slider control that can automatically color the points based on a k-means clustering.

We described a motivating challenge posed to us by an industrial partner in the financial services compliance industry. This involved the identification of suspicious patterns. We generated two synthetic datasets, using generators provided by our industrial partner, to illustrate the strengths and limitations of `EgoNav` as it pertains to the identification of exceptional egos. We also experimented with two real world datasets derived from a mobile phone call network and a peer-to-peer lending service. Using the egocentric spatialization, two analysts from our industrial partner in the financial service compliance industry were able to identify clusters of typical egos and a number of exceptional ones.

There are many directions for future work. At the end of Sect. 4.2 we noted the possibility of weighting the individual values in the network ratio profiles. This would impact the projection of the network ratio profiles onto points in the egocentric spatialization. For example, a user may choose to ignore the contribution of one network motif entirely or emphasize the contribution of another.

Finally, from a scalability perspective, the major bottleneck in our methodology is the network motif enumeration. We are investigating methods of speeding up this computation through parallelizing the enumeration for each egocentric network and using existing specialized algorithms to enumerate problematic network motifs.

6. REFERENCES

[1] L. Antiqueira and L. da Fontoura Costa. Characterization of Subgraph Relationships and Distribution in Complex Networks. *New Journal of Physics*, 11(013058), 2009.

[2] A. Bezerianos, F. Chevalier, P. Dragicevic, N. Elmqvist, and J. Fekete. GraphDice: A System for Exploring Multivariate Social Networks. *Computer Graphics Forum*, 29(3):863–872, 2010.

[3] V. Blondel, A. Gajardo, M. Heymans, P. Senellart, and P. Van Dooren. A Measure of Similarity between Graph Vertices. *SIAM Review*, 46(4):647–666, 2004.

[4] S. Borgatti and M. Everett. The Class of All Regular Equivalences: Algebraic Structure and Computation. *Social Networks*, 11:65–88, 1989.

[5] U. Brandes, J. Lerner, M. Lubbers, C. McCarty, and J. Molina. Visual Statistics for Collections of Clustered Graphs. In *Proc. of the IEEE VGTC Pacific Visualization Symp. (PacificVis'08)*, pages 47–54, 2008.

[6] U. Brandes, J. Lerner, U. Nagel, and B. Nick. Structural Trends in Network Ensembles. In *Proc. of the 1st Int'l. Workshop on Complex Networks (CompleNet'09)*, pages 83–97, 2009.

[7] N. Eagle, A. Pentland, and D. Lazer. Inferring Friendship

Network Structure by Using Mobile Phone Data. *PNAS*, 106(36):15274–15278, 2009.

[8] P. Erdös and A. Rényi. On Random Graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[9] M. Freire, C. Plaisant, B. Shneiderman, and J. Golbeck. ManyNets: An Interface for Multiple Network Analysis and Visualization. In *Proc. of the ACM Conf. on Human Factors in Computing Systems (CHI'10)*, pages 213–222, 2010.

[10] R. Ginoza and A. Mugler. Network Motifs Come in Sets: Correlations in the Randomization Process. *Physical Review E*, 82(1), 2010.

[11] R. Giugno and D. Shasha. GraphGrep: A Fast and Universal Method for Querying Graphs. In *Proc. of the 16th Int'l. Conf. on Pattern Recognition (ICPR'02)*, pages 112–115, 2002.

[12] W. Huang, C. Murray, X. Shen, L. Song, Y. Xin Wu, and L. Zheng. Visualisation and Analysis of Network Motifs. In *Proc. of the 9th Int'l. Conf. on Information Visualisation (IV'05)*, pages 697–702, 2005.

[13] D. Jeong, C. Ziemkiewicz, B. Fisher, W. Ribarsky, and R. Chang. iPCA: An Interactive System for PCA-Based Visual Analytics. In *Proc. of the 11th Eurographics/IEEE Symp. on Visualization (EuroVis'09)*, pages 767–774, 2009.

[14] R. Khan, M. Corney, A. Clark, and G. Mohay. Transaction Mining for Fraud Detection in ERP Systems. *Industrial Engineering & Management Systems*, 9(2):141–156, 2010.

[15] C. Klukas, F. Schreiber, and H. Schwöbbermeyer. Coordinated Perspectives and Enhanced Force-Directed Layout for the Analysis of Network Motifs. In *Proc. of the 5th Asia-Pacific Symp. on Visualization (APVIS'06)*, pages 39–48, 2006.

[16] D. Koschützki, H. Schwöbbermeyer, and F. Schreiber. Ranking of Network Elements Based on Functional Substructures. *Journal of Theoretical Biology*, 248(3):471–479, 2007.

[17] E. Leicht, P. Holme, and M. Newman. Vertex Similarity in Networks. *Physical Review E*, 73(026120), 2006.

[18] C. Li and S. Lin. Egocentric Information Abstraction for Heterogeneous Social Networks. In *Proc. of the 1st Int'l. Conf. on Social Networks Analysis & Mining (ASONAM'09)*, pages 255–260, 2009.

[19] F. Lorrain and H. White. Structural Equivalence of Individuals in Social Networks. *Journal of Mathematical Sociology*, 1:49–80, 1971.

[20] M. Lubbers, J. Molina, J. Lerner, U. Brandes, J. Ávila, and C. McCarty. Longitudinal Analysis of Personal Networks: The Case of Argentinean Migrants in Spain. *Social Networks*, 32(1):91–104, 2010.

[21] A. Ma'ayan, S. Jenkins, R. Webb, S. Berger, S. Purushothaman, N. Abul-Husn, J. Posner, T. Flores, and R. Iyengar. SNAVI: Desktop Application for Analysis and Visualization of Large-Scale Signaling Networks. *BMC Systems Biology*, 3(10), 2009.

[22] B. McKay. Isomorph-Free Exhaustive Generation. *Journal of Algorithms*, 26(2):306–324, 1998.

[23] F. McSherry. Spectral Partitioning of Random Graphs. In *Proc. of the 42nd Annual Symp. on Foundations of Computer Science (FOCS'01)*, pages 529–537, 2001.

[24] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of Evolved and Designed Networks. *Science*, 303(5663):1538–1542, 2004.

[25] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594):824–827, 2002.

[26] S. Roweis and L. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 22(5500):2323–2326, 2000.

[27] F. Schreiber and H. Schwöbbermeyer. MAVisto: A Tool for the Exploration of Network Motifs. *Bioinformatics*, 21(17):3572–3574, 2005.

[28] A. Stoica and C. Prieur. Structure of Neighborhoods in a Large Social network. In *Proc. of the Int'l. Conf. on Computational Science & Engineering (CSE'09)*, pages 26–33, 2009.

[29] J. Tenenbaum, V. de Silva, and J. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 22(290):2319–2323, 2000.

[30] M. Tory, D. Sprague, F. Wu, W. So, and T. Munzner. Spatialization Design: Comparing Points and Landscapes. *IEEE Trans. on Visualization & Computer Graphics (TVCG)*, 13(6):1262–1269, 2007.

[31] T. von Landesberger, M. Görner, and T. Schreck. Visual Analysis of Graphs with Multiple Connected Components. In *Proc. of the IEEE Symp. on Visual Analytics Science & Technology (VAST'09)*, pages 155–162, 2009.

[32] D. Watts and S. Strogatz. Collective Dynamics of 'Small-World' Networks. *Nature*, 393:440–442, 1998.

[33] B. Wellman. An Egocentric Network Tale: Comment on Bien et al. *Social Networks*, 15:423–436, 1993.

[34] H. Welser, E. Gleave, D. Fisher, and M. Smith. Visualizing the Signatures of Social Roles in Online Discussion Groups. *Journal of Social Structure*, 8, 2007.

[35] S. Wernicke and F. Rasche. FANMOD: A Tool for Fast Network Motif Detection. *Bioinformatics*, 22(9):1152–1153, 2006.

[36] H. White, S. Boorman, and R. Breiger. Social Structure from Multiple Networks – Blockmodels of Roles and Positions. *American Journal of Sociology*, 81:730–780, 1976.