



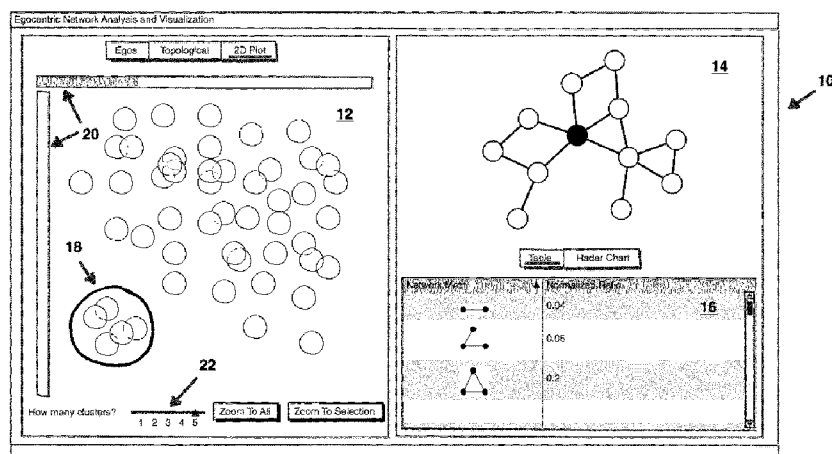
- (51) **International Patent Classification:**
H04L 29/08 (2006.01) *G06F 19/12* (2011.01)
- (21) **International Application Number:**
PCT/EP2012/056303
- (22) **International Filing Date:**
5 April 2012 (05.04.2012)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
1107251.9 3 May 2011 (03.05.2011) GB
- (71) **Applicant (for all designated States except US):** UNIVERSITY COLLEGE DUBLIN, NATIONAL UNIVERSITY OF IRELAND, DUBLIN [IE/IE]; Belfield, Dublin 4 (IE).
- (72) **Inventors; and**
- (75) **Inventors/Applicants (for US only):** HARRIGAN, Martin [IE/IE]; Harrigan, Clonmel (IE). ARCHAMBAULT, Daniel [CA/IE]; H6A9 Sandford Lodge Apts, Sandford Close, Ranelagh, Dublin 6 (IE).
- (74) **Agent:** BOYCE, Conor; 27 Clyde Road, Dublin 4 (IE).

(81) **Designated States** (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

(54) **Title:** NETWORK ANALYSIS TOOL**Figure 1**

(57) **Abstract:** A network analysis tool for analyzing a network of information comprising a plurality of interconnected nodes is disclosed. The tool determines, for the network, a set of network motifs, each network motif comprising a respective pattern of connections between a node and at least its neighbouring nodes. For each node, a network motif profile is determined, the profile comprising for each network motif of the set of network motifs, a count of the instances of the network motif at the node. The network motif profile for each node is normalized relative to the network motif profiles of other nodes in the network. The normalized network motif profiles for the network are projected from a high-dimensional space corresponding to the number of motifs in the set of network motifs, onto a lower dimensional space based on maximizing the variability of normalized network motif profiles with the space. At least some of the nodes of the network are displayed in the lower dimensional space.

Network Analysis Tool

Field of the Invention

The present invention relates to a method and tool for analyzing a network of
5 information

Background

There are many examples of networked information in data processing. For example,
in social networks, network nodes correspond with people and connections between
10 nodes can represent relationships between friends and family. In financial networks,
the affairs of payees and payers, or borrowers and lenders can be represented by
account nodes interconnected by transactions between those nodes. Similarly in
telephone or communications networks, nodes corresponding to telephone or e-mail
accounts can be related by calls, messages or correspondence between those
15 accounts.

The network of connections around a node of a network can provide characteristic
information about the node itself.

20 It is appreciated in that art that the structural similarity of two nodes within a network
can be measured in two ways. LORRAIN F., WHITE H.: Structural Equivalence of
Individuals in Social Networks, Journal of Mathematical Sociology 1 (1971), 49–80
disclose that two nodes are structurally equivalent if they share many of the same
neighbors. LEICHT E., HOLME P., NEWMAN M.: Vertex Similarity in Networks,
25 Physical Review E 73, 026120 (2006) formulate a measure based on two nodes
being regularly equivalent if they are connected to other nodes that are themselves
structurally similar.

Structural similarity can also be used to classify entire networks. BRANDES U.,
30 LERNER J., NAGEL U., NICK B.: Structural Trends in Network Ensembles,
Proceedings of the 1st International Workshop on Complex Networks (CompleNet'
09) (2009), Springer, pp. 83–97 discloses an effective heuristic for partitioning a set
of (planted partition) networks such that two networks are in the same part if and only

if they are generated using the same random network model. The 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.

5 SHEN-ORR S., MILO R., MANGAN S., ALON U.: Network Motifs in the Transcriptional Regulation Network of Escherichia Coli. *Nature Genetics* 31 (2002), 64–68; and MILO R., SHEN-ORR S., ITZKOVITZ S., KASHTAN N., CHKLOVSKII D., ALON U.: Network Motifs: Simple Building Blocks of Complex Networks, *Science* 298, 5594 (2002), 824–827 disclose forms of network motif analysis. This represents
10 the structural or topological properties of a network by counting all occurrences of each network motif in a network.

MILO R., ITZKOVITZ S., KASHTAN N., LEVITT R., SHEN-ORR S., AYZENSHTAT I., SHEFFER M., ALON U.: Superfamilies of Evolved and Designed Networks, *Science*
15 303, 5663 (2004), 1538–1542 use network ratio profiles to compare the counts of network motifs in 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, Internet networks, social networks and autonomous systems networks.

20 KOSCHÜTZKI D., SCHWÖBBERMEYER H., SCHREIBER F.: Ranking of Network Elements Based on Functional Substructures, *Journal of Theoretical Biology* 248, 3 (2007), 471–479 formulate a number of network motif-based centrality measures. They rank the vertices of the E. Coli transcriptional network using each centrality
25 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
30 counts.

There are also several visualization systems that aid with network motif analysis, such as disclosed in SCHREIBER F., SCHWÖBBERMEYER H.: MAVisto: A Tool for

the Exploration of Network Motifs, *Bioinformatics* 21, 17 (2005), 3572–3574; and MA'AYAN A., JENKINS S., WEBB R., BERGER S., PURUSHOTHAMAN S., ABUL-HUSN N., POSNER J., FLORES T., IYENGAR R.: SNAVI: Desktop Application for Analysis and Visualization of Large-Scale Signaling Networks, *BMC Systems Biology* 3, 10 (2009). However, these are concerned with network motif analysis on entire networks.

WHITE H., BOORMAN S., BREIGER R.: Social Structure from Multiple Networks – Blockmodels of Roles and Positions, *American Journal of Sociology* 81 (1976), 730–780; BORGATTI S., EVERETT M.: The Class of All Regular Equivalences: Algebraic Structure and Computation, *Social Networks* 11 (1989), 65–88; and WELLMAN B.: An Egocentric Network Tale: Comment on Bien et al, *Social Networks* 15 (1993), 423–436 each disclose node based analysis within the field of social network analysis.

LUBBERS M., MOLINA J., LERNER J., BRANDES U., ÁVILA J., MCCARTY C.: Longitudinal Analysis of Personal Networks: The Case of Argentinean Migrants in Spain, *Social Networks* 32, 1 (2010), 91–104 describe a dynamic node based network analysis of Argentinean immigrants in Spain. The analysis comprises qualitative interviews and a quantitative analysis at three distinct levels (node-neighbour pairs, neighbour-neighbour pairs and networks). The quantitative analysis investigated the characteristics of the nodes, the structural characteristics of the networks, the characteristics of the node-neighbour pairs and neighbours, the structural positions of the neighbours, and the characteristics of the neighbour-neighbour pairs.

In BRANDES U., LERNER J., LUBBERS M., MCCARTY C., MOLINA J.: Visual Statistics for Collections of Clustered Graphs, In *Proceedings of the IEEE VGTC Pacific Visualization Symposium (PacificVis'08)* (2008), IEEE, pp. 47–54, the composition of networks were visualized using clustered networks 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 H., GLEAVE E., FISHER D., SMITH M.: Visualizing the Signatures of Social Roles in Online Discussion Groups, *Journal of Social Structure* 8 (2007) present an analysis of roles in an online discussion group. They visualize posting habits within networks where through visual inspection, they identify three types of poster: answer people, discussion people, and disruptors.

Similarly, STOICA A., PRIEUR C.: Structure of Neighborhoods in a Large Social network, *Proceedings of the International Conference on Computational Science and Engineering (CSE'09)* (2009), pp. 26–33 analyze a large mobile phone call network and partition nodes according to roles and validate their results using network attributes.

ANTIQUEIRA L., DA FONTOURA COSTA L.: Characterization of Subgraph Relationships and Distribution in Complex Networks, *New Journal of Physics* 11, 013058 (2009) present a methodology for analyzing non-overlapping subnetworks, their interrelationships, and their distribution in a network. Given a network and a set of non-overlapping subnetworks, they generate histograms of the subnetwork sizes and the shortest distances between the subnetworks. They then dilate each subnetwork, merging subnetworks when necessary, until the entire network is covered. They consider the rate at which the subnetworks merge and the rate at which vertices are covered by the dilations. 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.

The analyses of Welser et al., Stoica and Prieur and Antiqueira and da Fontoura Costa choose a number of network statistics, for example, degree, clustering coefficient and local triangle count, when characterizing subnetworks and the choice of network statistics is often specific to the task at hand.

Although not directly related to node based visualization, VON LANDESBERGER T., GÖRNER M., SCHRECK T.: Visual Analysis of Graphs with Multiple Connected Components, *Proceedings of the IEEE Symposium on Visual Analytics Science and*

Technology (VAST'09) (2009), IEEE Computer Society, pp. 155–162 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, including reciprocity features, distance features, clustering features and degree distribution features. The user weights the features appropriately and the system produces a SOM layout.

Each cell represents a subset of similar networks; the background color indicates the number. Also, JEONG D., ZIEMKIEWICZ C., FISHER B., RIBARSKY W., CHANG R.: iPCA: An Interactive System for PCA-Based Visual Analytics, Proceedings of the 11th Eurographics/IEEE Symposium on Visualization (EuroVis'09) (2009), pp. 767–774 discloses an interactive interface that opens up the black box of Principle Component Analysis (PCA).

Given the above, it will be appreciated that current network analysis and visualization tools focus on analyzing and visualizing either the entire network or individual node networks but fail to visually summarize a collection of node networks.

Other approaches include E-NET, which is a tool developed primarily by the social scientist Steve Borgatti, for analyzing networks. It aids with data collection, data analysis and visualization. For data collection, it produces appropriate questionnaires to elicit attribute and relationship data from people. For data analysis, it measures size, composition (e.g. homogeneity and homophily), structure (e.g. connectedness, density and structural holes), etc. For visualization, it produces tabular representations of the results.

LI C., LIN S.: Egocentric Information Abstraction for Heterogeneous Social Networks. In Proceedings of the 1st International Conference on Social Networks Analysis and Mining (ASONAM'09) (2009), IEEE Computer Society, pp. 255–260 summarize egocentric networks by combining the surrounding relational structures with the statistical dependencies between attribute values to form feature vectors. The features describing the relational structures are based on the various types of paths of fixed length (say, two) that can emanate from an ego (node). They use frequency-based measures (local frequency, local rarity and relative frequency) to determine

whether or not a feature is relevant. They then construct representative egocentric networks using only the relevant features.

In spite of these approaches, there remains a need for tools and techniques for effectively analyzing and visually summarizing networks.

Summary

According to the present invention there is provided a computer-implemented method for analyzing a network of information comprising a plurality of interconnected nodes, the method comprising the steps of:

for the network, determining a set of network motifs, each network motif comprising a respective pattern of connections between a node and at least its neighbouring nodes;

for each node, determining a network motif profile comprising for each network motif of the set of network motifs, a count of the instances of the network motif at said node;

for each node, normalizing the network motif profile relative to the network motif profiles of other nodes in the network;

for the network, projecting the normalized network motif profiles from a high-dimensional space corresponding to the number of motifs in said set of network motifs, onto a lower dimensional space based on maximizing the variability of normalized network motif profiles with said space; and displaying at least some of the nodes of said network in said lower dimensional space.

Preferably, said dimensionality reduction comprises performing principal component analysis (PCA) on said normalized network motif profile information in said high dimensionality space.

The present invention provides a system that analyzes and clusters nodes based on the relationship structure of their network connections; and presents the results as a node based spatialization.

Embodiments of the invention use a form of network motif analysis and dimensionality reduction to cluster nodes so that two nodes are in the same cluster if their respective network connections are structurally similar. This view of a network discriminates between the various classes of typical and exceptional nodes.

5

Embodiments of the present invention combine network motif analysis at the node level and dimensionality reduction using PCA to produce an aggregated node based view of a network. Embodiments allow a user to visually inspect networks, network ratio profiles, and a spatialization of the nodes based on the structural similarity of the node networks. The various views are coordinated allowing a user to select a node in one view and examine its properties in another. A user can also compare, for example, network ratio profiles through selecting multiple nodes to help identify the distinguishing features of a collection of node networks.

10

15

Embodiments of the invention use network motif analysis to exhaustively count the number of network motifs up to a certain size in a network. For large networks, this could be prohibitive, but for a collection of node networks, the computation can be divided and parallelized.

20

A node's network connections can include connections between a node and its immediate neighbours as well as connections between a node's neighbours and possibly their neighbours.

25

Embodiments of the invention are particularly useful for identifying rogue behaviour without *a priori* knowledge of the form of this behaviour. For example, if a personal bank account in a financial transaction network is typical, its network connections should be structurally similar to network connections of other typical accounts. At the very least, there should be a small number of classes of typical accounts. On the other hand, if a bank account is involved in smurfing (the splitting of large financial transactions into multiple smaller transactions, each of which is below a limit above which financial institutions must report), assuming the incidence of smurfing is relatively low, the bank account's network connections should be relatively

30

exceptional. The only inputs for required for the present system to analyze such a network would be a list of account transactions.

In embodiments of the invention, the structure of a node's network is defined by the longest shortest-path distance k from a node to every other node in the node's network (the radius) as well as the various network motifs to be counted. Preferably, the counts for each node's network are adjusted for scale to produce network ratio profiles. The network ratio profiles can be interpreted as points in a high-dimensional space. Preferably, they are projected onto a 2-dimensional spatialization using principal component analysis (PCA). This projection removes the correlations between the counts. The spatialization encodes the similarities and differences amongst the node networks. Furthermore, clusters of nodes represent broad classes of nodes with structurally similar node networks.

Brief Description of the Drawings

Embodiments of the invention will now be described, by way of example, with reference to the accompanying drawings, in which:

Figure 1 shows a user interface including a view generated according to an embodiment of the present invention for browsing a single 1,000-node random network from the ER dataset.

Figure 2 shows a summary for a selection of nodes in Figure 1.

Figure 3 shows a view of a single 1,000-node network from the WS dataset generated according to an embodiment of the present invention.

Figure 4 shows a view of the activity in the Prosper Marketplace dataset during April 2010 generated according to an embodiment of the present invention.

Figure 5 shows a comparison of a view generated according to an embodiment of the present invention and a global view of the MIT Reality Mining dataset.

Figure 6 is a flow diagram illustrating an embodiment of the invention.

Description of the Preferred Embodiments

Embodiments of the present invention comprise a network analysis tool which produces a spatialization for a network of nodes (egos) that clusters nodes so that two nodes are in the same cluster if their node networks are structurally similar. One of the potential applications for the tool is in identifying and visualizing nodes exhibiting potentially fraudulent behaviour without specifying the behaviour of such nodes *a priori*.

Referring now to Figure 1, which shows a user interface 10 for a network analysis tool according to an embodiment of the present invention. The interface comprises multiple coordinated views 12, 14, 16. Each of the three views 12, 14, 16 illustrates a specific aspect of selected node(s) and each view is coordinated with the others, so that selecting nodes in one window causes appropriate updates in the other views. The views are coordinated, allowing a user to select a node in one view and examine its properties in the others. For example, the system includes a view 14 of the topology of the selected node networks and a view 16 for comparing their network ratio profiles. The spatialization 12 is the central view in the system. A user may pan and zoom within this view. At the top left, the view 12 includes bar indicators 20. The length of each bar 20 shows the percentages of variability captured by each axis of the view; as such, these can be interpreted as a measure of the significance of each axis of the spatialization. The view 12 further includes a slider control 22 that can automatically color the nodes based on a k-means clustering.

The node based spatialization in the view 12 is computed through network motif analysis and dimensionality reduction, described in more detail below.

Referring now to Figure 6, the tool begins by calculating a node network for each node in turn, step 60. The node network, or k-neighborhood subnetwork, of a node u is the subnetwork induced by the set of vertices that have shortest-path distance at most k hops from u . In the illustrated embodiment, $k = 2$, i.e. a node's network extends to its neighbours and its neighbours' neighbours.

For each node network, a network motif profile is calculated, step 62. In the embodiment, a profile comprises a 30-element vector where each entry is based on a count of the number of instances of the corresponding network motif in an ordered list that are incident with the node.

The ordered list comprises all network motifs with at most l vertices up to isomorphism connected to a node. In the illustrated embodiment, $l = 5$, i.e. the maximum number of nodes in any given network motif is 5. The ordered list can be generated using, for example, geng from the nauty package disclosed in MCKAY B.: Isomorph-Free Exhaustive Generation. Journal of Algorithms 26, 2 (1998), 306–324.

The counts for each element of the network profile vector can be calculated using GraphGrepSX disclosed in GIUGNO R., SHASHA D.: GraphGrep: A Fast and Universal Method for Querying Graphs, Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02) (2002), pp. 112–115. GraphGrepSX is a tool that solves the subgraph isomorphism problem using enumerated paths as index features. This can be a time-consuming process, but for large datasets, node networks can be processed in parallel and/or both k and l can be reduced.

For each network motif profile, a network ratio profile is computed, step 64. In the network ratio profile, each entry of the 30-element vector comprises a normalized ratio of the corresponding entry in the network motif profile. The ratio profile rp of a node 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 node networks. To adjust for scaling, the normalized ratio profile nrp of a node network is computed using:

$$nrp_i = \frac{rp_i}{(\sum rp_i^2)^{1/2}}$$

A normalized ratio measures the abundance of a network motif in each individual node network relative to all node networks; it is similar to a z-score.

5 It is noted that there are correlations between the elements of a network ratio profile. Thus, in the embodiment, to adjust for these, a dimensionality reduction is performed, step 66. Principle Component Analysis (PCA) is an exemplary dimensionality reduction technique that calculates the eigenvectors of a covariance matrix generated from a set of vectors, in this case, the network ratio profiles. Each
10 eigenvector, or principal component, corresponds to an orthogonal direction of variation within the data. Often, a small subset of eigenvectors can account for much of the variability. PCA identifies underlying structures such as clusters and outliers that are difficult to perceive in the original set of vectors. In the present embodiment, the first two eigenvectors are calculated and these are used as the x and y axes for
15 the spatialization shown in the window 12 of Figure 1. It will of course be appreciated that if more than two eigenvectors were calculated, the window 12 could be implemented as a three dimensional display of nodes; or other techniques for enabling browsing of higher dimension spaces could be employed. It will also be noted that in Figure 1, the PCA analysis has been directed to determine 5 clusters of
20 nodes, whereas it can be seen that any number of clusters can be chosen.

Figure 1 is based on an analysis of the ER dataset. This includes 50 random networks generated using the Erdős-Rényi model disclosed in: ERDÖS P., RÉNYI A.: On Random Graphs. Publicationes Mathematicae Debrecen 6 (1959), 290–297;
25 and ERDÖS P., RÉNYI A.: On the Evolution of Random Graphs. Institute of Mathematics, Hungarian Academy of Sciences 5 (1960), 17–61. [GM10] GINOZA R., MUGLER A.: Network Motifs Come in Sets: Correlations in the Randomization Process, Physical Review E 82, 1 (2010). Each network contains 1,000 vertices and 5,000 edges. Furthermore, five nodes were chosen at random and augmented with
30 additional edges to create a clique.

In Figure 1, a black circle 18 indicates the five nodes belonging to the five-node clique. The exceptionality of these egos is apparent from their position within the window 12. When one of the nodes is selected, the node network for the selected node is displayed in the window 14, while at the same time the network ratio profile for the selected node is displayed in the window 16. The normalized ratio ranges from [-1,1] and so it will be seen that at least for the first six elements of the vector the profile is relatively close to average.

Figure 2 shows an updated window 16 from Figure 1, if all five vertices of the clique indicated by the circle 18 in the network in Fig. 1 are selected and displayed in radar chart style. Figure 2 shows the five nodes v0...v4 have broadly identical network ratio profiles making them structurally similar. All elements in the network ratio profiles are relatively high, especially for the higher-order network motifs and this is what makes the corresponding nodes exceptional.

The WS dataset also includes 50 random networks generated using the Watts-Stogatz model disclosed in WATTS D., STROGATZ S.: Collective Dynamics of 'Small-World' Networks. Nature 393 (1998), 440–442. Again, each network contains 1,000 vertices and 5,000 edges. Again, five vertices (nodes) were chosen at random and augmented with additional edges to create a clique.

The ER and WS datasets reveal both the strength and limitation of the present approach. In Figure 1 based on the ER dataset, the clique comprising nodes v0...v4 is easily identifiable through the spatialization 12. (The same is true for the other networks in the dataset.) These cliques are not easily identifiable in the corresponding topological views produced by a force-directed algorithm, for example, the view 26 shown in Figure 5.

However, Figure 3, which is based on a single 1,000-node network from the WS dataset is less convincing. The five nodes belonging to the five node clique are surrounded by a black circle 33. The clique is not easily identifiable through the spatialization shown. This is due to the increased clustering coefficient (indicated on the axis 32) found in networks from the WS dataset compared to those from the ER

dataset. The nodes in the clique are no longer considered exceptional. (Again, the same is also true for all 50 networks in the WS dataset). The networks in both Fig. 1 and Fig. 3 have the same number of vertices and edges. However, their differing structure means that a node centric network considered exceptional in one is typical in another.

The Prosper Marketplace dataset (www.prosper.com) is derived from a peer-to-peer lending or social lending service where 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 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 of a loan agreed upon between a borrower and a lender. It should be noted that lenders can also be borrowers and vice versa and therefore the network is not necessarily bipartite.

Figure 4 includes a view showing 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 42) account for 54% and 16% of the variability in the original dataset (see the bar indicators 44). It can be seen that the node networks of nodes to the left of the spatialization have more vertices and edges than those of the nodes to the right. However, the difference between the nodes along the y-axis of the spatialization is more interesting. Two representative nodes 46 are selected in Fig. 4. The radar chart view 48 of their network ratio profiles reveal that the node to the top, when compared to the node to the bottom, has a node network with relatively fewer lower-order network motifs but relatively more higher-order network motifs. This is corroborated by the small multiples representation 49 of the two corresponding node networks. The node to the bottom of the spatialization (to the left of the small multiples representation 49) has just two neighbours, both of whom are connected to many others. The vertices at the center

of the two circles 51 represent the two neighbours. However, the node to the top of the spatialization (to the right of the small multiples representation) has many more neighbours. The vertices in the circle surrounding the node represent these.

- 5 The differences between, say, the top and bottom egos in Fig. 4 can be computed more easily and directly using, say, node degrees and clustering coefficients. However, the importance and flexibility of the above approach lies in the fact that the nature or the distinguishing feature(s) of the differences was not input *a priori*. Thus, through the exploration of the spatialization 42 and the network ratio profiles 48 we
10 can deduce the distinguishing feature(s) of nodes in cliques.

The MIT Reality Mining dataset disclosed in EAGLE N., PENTLAND A., LAZER D.: Inferring Friendship Network Structure by Using Mobile Phone Data, Proceedings of the National Academy of Sciences (PNAS) 106, 36 (2009), 15274–15278 comprises
15 mobile phone call and SMS records over a 296-day period between 100 unique mobile phones. The dataset is 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. In this case, a node represents a user, or more specifically a mobile phone, and an edge represents a mobile phone
20 call or SMS between two mobile phones. Figure 5 shows a node based view 12' produced according to an embodiment of the present invention and a global view 26 produced using a force-directed algorithm of the network indicating all calls between all users. The global view 26 identifies two large communities, a known artifact of the dataset, being mobile phone users with dense communication within each group and
25 sparse communication between the groups.

The view 12' also identifies two communities 30', 30" but these do not correspond to the two communities in the global view 26. Instead, they correspond to core mobile phone users 30" and peripheral mobile phone users 30'. The peripheral mobile
30 phone users can be further divided into an inner periphery (the nodes 30A below the divider) and an outer periphery (the nodes 30B above the divider). The selected nodes within the circle 30' in the view 12' correspond to the selected nodes in the two circles 28', 28" in the global view 26.

In extensions of the embodiments described above, specialized algorithms could be employed to enumerate more complex network motifs, for example, stars and triangles.

5

It could also be possible to allow for weighting individual elements of the network ratio profile vector. This would effect the projection of the network ratio profiles onto points in the spatialization. For example, a user could choose to ignore the contribution of one network motif entirely or emphasize the contribution of another.

10

It will be seen that in some cases, using PCA, the percentage of variability accounted for by either of the axes in the view 12 could be relatively small, as would be indicated by short bar indicators, and so the positioning of the nodes along the axes would not be significant. Using other dimensionality reduction techniques such as disclosed in ROWEIS S., SAUL L.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 22, 5500 (2000), 2323–2326 could preserve locality and allow for better presentation of data in these cases.

15

Claims:

1. A computer-implemented method for analyzing a network of information comprising a plurality of interconnected nodes, the method comprising the steps of:
5 for the network, determining a set of network motifs, each network motif comprising a respective pattern of connections between a node and at least its neighbouring nodes;
for each node, determining a network motif profile comprising for each network motif of the set of network motifs, a count of the instances of the network motif at said
10 node;
for each node, normalizing the network motif profile relative to the network motif profiles of other nodes in the network;
for the network, projecting the normalized network motif profiles from a high-dimensional space corresponding to the number of motifs in said set of network
15 motifs, onto a lower dimensional space based on maximizing the variability of normalized network motif profiles with said space; and
displaying at least some of the nodes of said network in said lower dimensional space.
- 20 2. The method of claim 1 wherein said projecting comprises performing principal component analysis (PCA) on said normalized network motif profile for said nodes.
3. The method of claim 1 further comprising the step of: responsive to a user selecting one or more nodes from said lower dimensional space display,
25 simultaneously displaying network motif profile information for said selected nodes.
4. The method of claim 1 further comprising the step of: responsive to a user selecting a node from said lower dimensional space display, simultaneously displaying a node network for said node.
30
5. The method of claim 1 wherein one or both of said determining and normalizing steps are performed in parallel for each node.

6. The method of claim 1 wherein a node's network connections includes connections between a node and its immediate neighbours and connections between a node's neighbours and their neighbours.

5

7. The method of claim 1 wherein said nodes correspond with bank accounts and said connections correspond with transactions between said bank accounts, said displaying enabling the identification *a posteriori* of potentially fraudulent behaviour between bank accounts.

10

8. The method of claim 1 wherein said nodes correspond with phone accounts and said connections correspond with connections between said phone accounts, said displaying enabling the identification *a posteriori* of irregular behaviour between users of said accounts.

15

9. The method of claim 1 wherein a node's network motifs extend a maximum number k of interconnections from a node to every other node in the node's network.

10. The method of claim 9 wherein $k=2$.

20

11. The method of claim 2, wherein said PCA analysis reduces said normalized network motif profile information to a 2-dimensional space.

12. The method of claim 1 further comprising the step of weighting one or more individual network motifs within said network motif profiles to either emphasize or de-emphasize variations between nodes in said network in respect of specific network motifs.

13. The method of claim 1 further comprising applying a second clustering to said network of nodes to divide said nodes into a specified number of clusters, and wherein said displaying comprises displaying said nodes in said lower dimensional space according to their designated cluster.

30

14. The method of claim 13 wherein said second clustering comprises k-means clustering.

15. A computer program product comprising computer readable instructions
5 stored on a computer readable medium which when executed in a computing device are arranged to perform the steps of any previous claim.

16. A network analysis tool for analyzing a network of information comprising a plurality of interconnected nodes, the tool being arranged to:

10 for the network, determine a set of network motifs, each network motif comprising a respective pattern of connections between a node and at least its neighbouring nodes;

for each node, determine a network motif profile comprising for each network motif of the set of network motifs, a count of the instances of the network motif at said node;

15 for each node, normalize the network motif profile relative to the network motif profiles of other nodes in the network;

for the network, project the normalized network motif profiles from a high-dimensional space corresponding to the number of motifs in said set of network motifs, onto a lower dimensional space based on maximizing the variability of normalized network

20 motif profiles with said space; and

display at least some of the nodes of said network in said lower dimensional space.

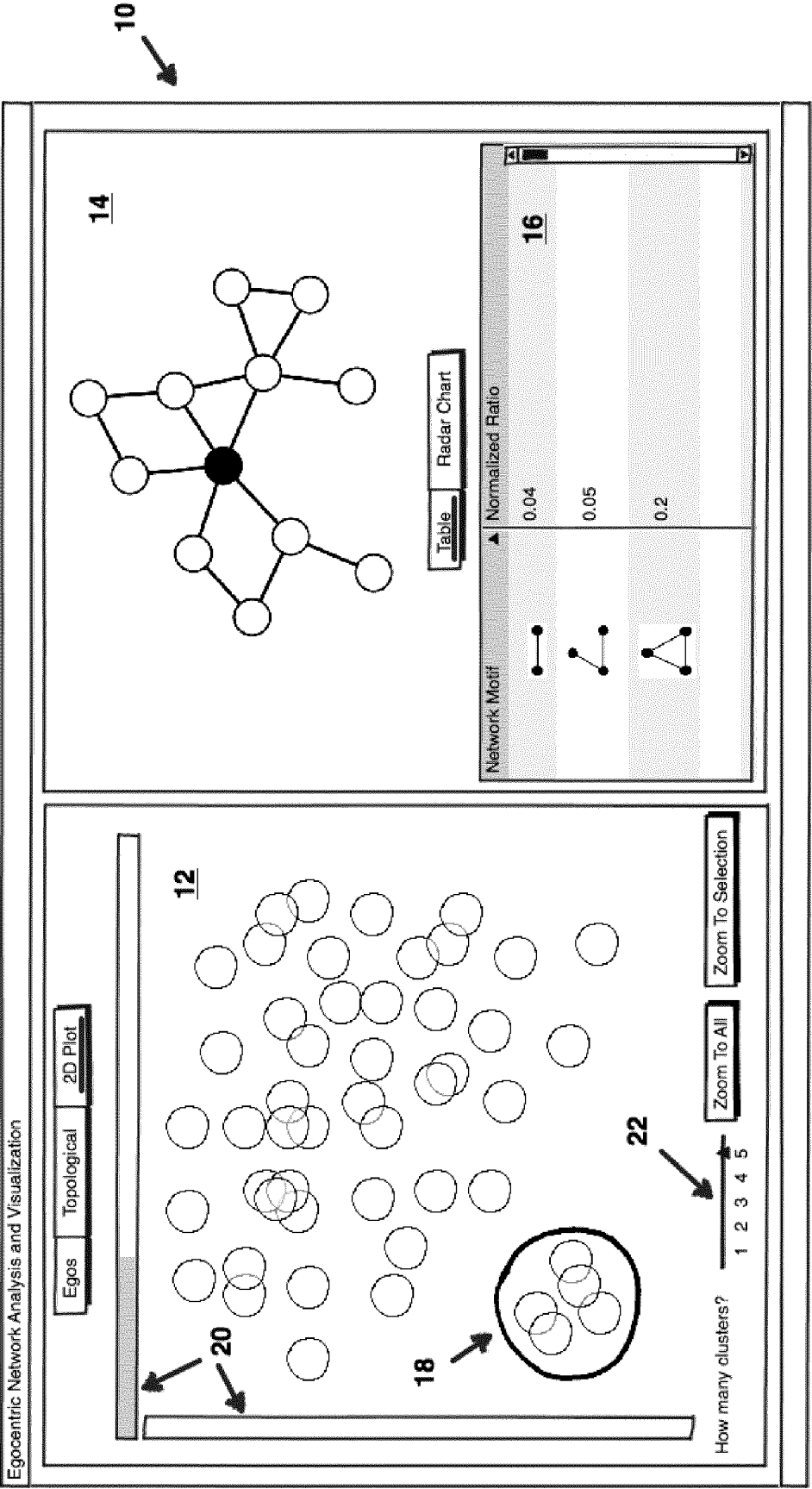


Figure 1

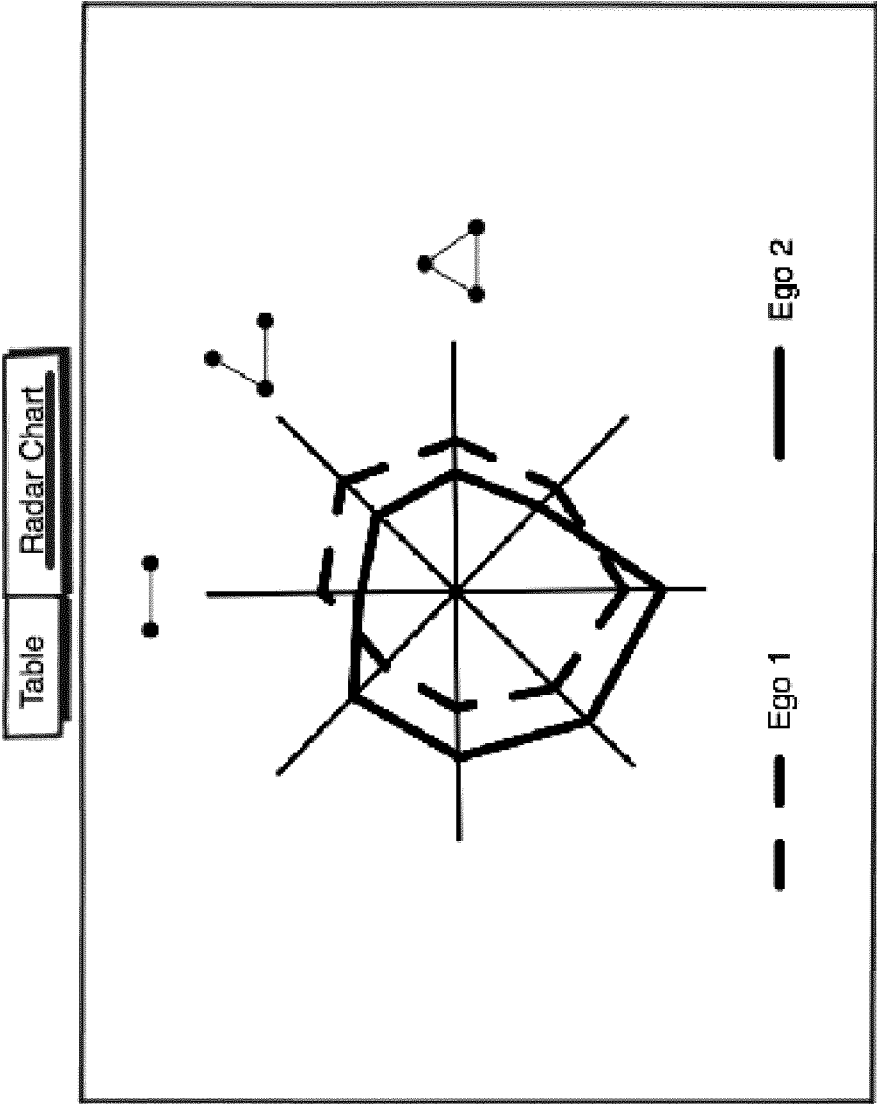
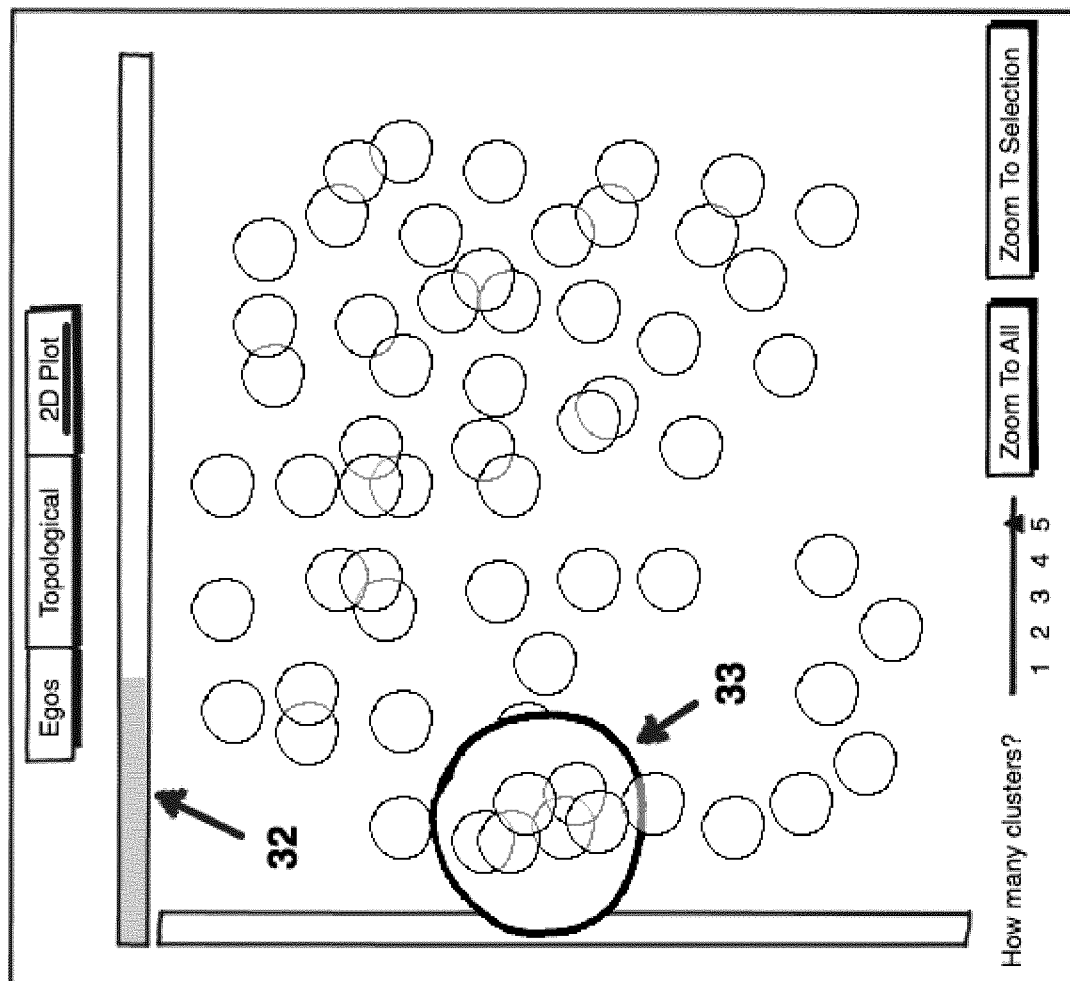


Figure 2

**Figure 3**

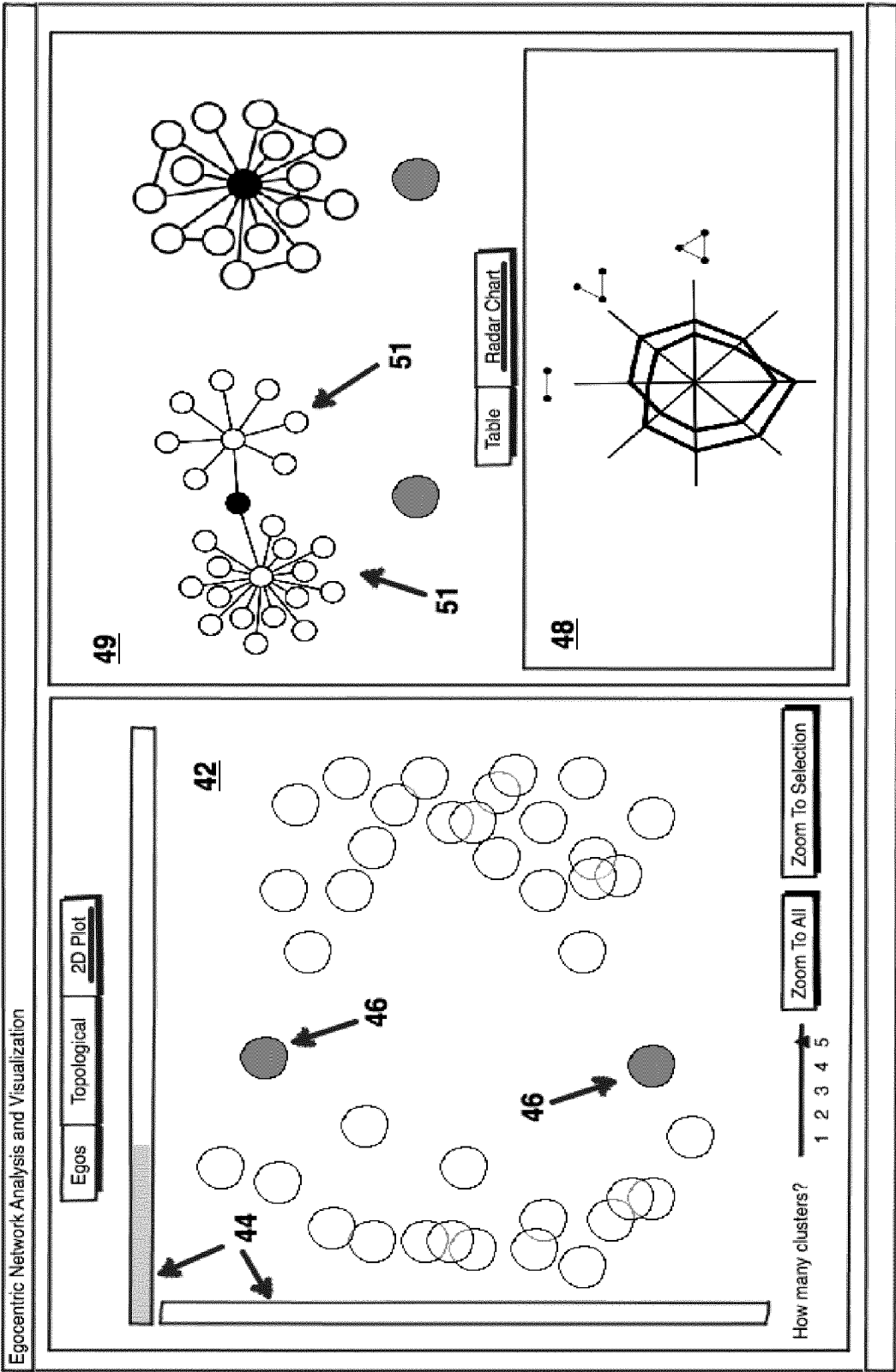


Figure 4

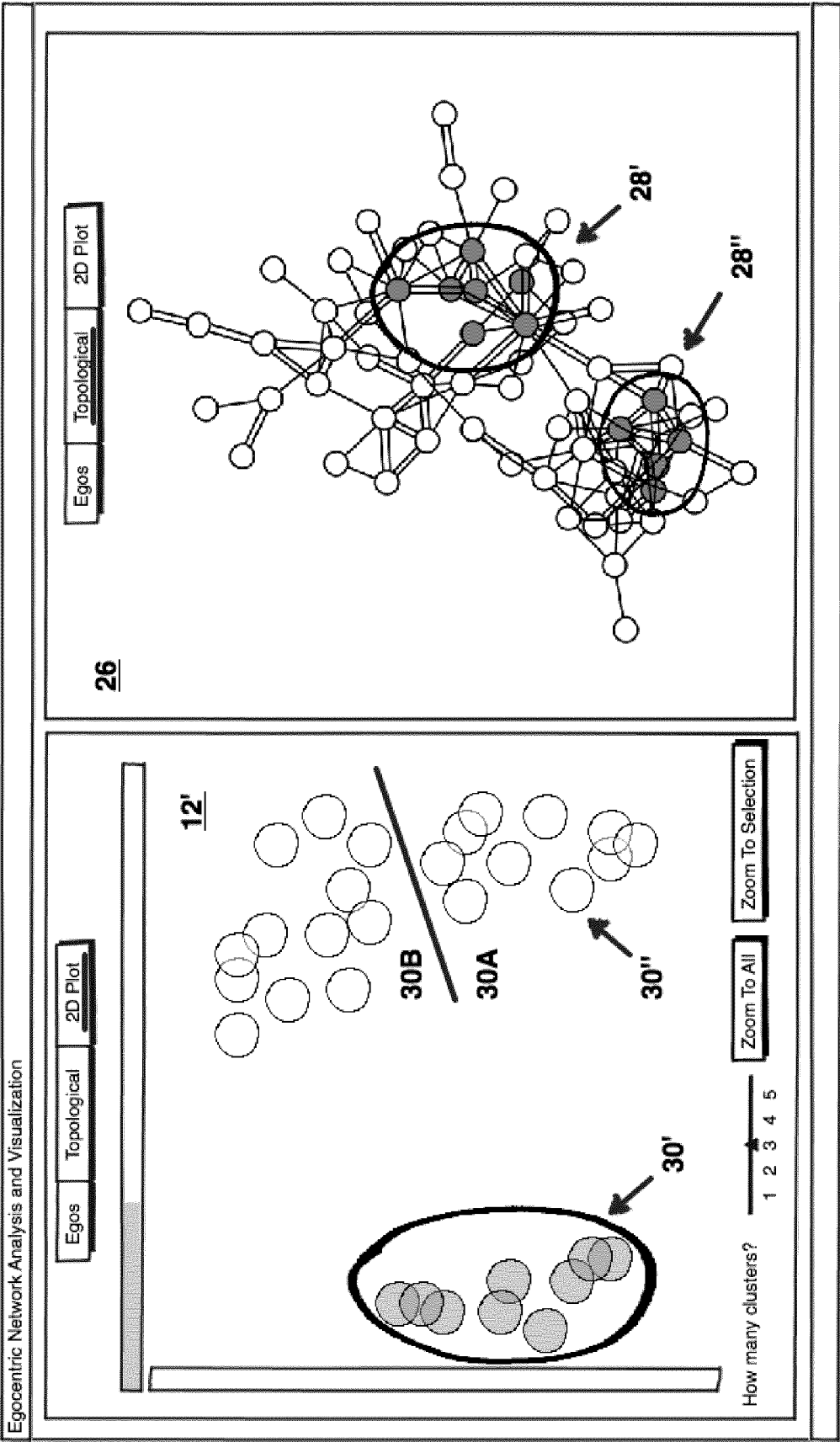


Figure 5

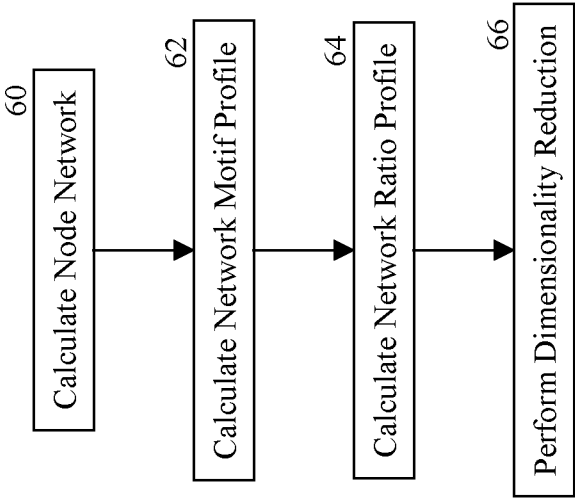


Figure 6

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2012/056303

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04L29/08 G06F19/12
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
H04L G06Q G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 03/062943 A2 (YEDA RES & DEV [IL]; ALON URI [IL]; SHEN-ORR SHAI S [IL]; MILO RON [IL]) 31 July 2003 (2003-07-31) page 2, line 17 - page 3, line 19 -----	1-16
X	US 2007/094066 A1 (KUMAR SHAI LESH [US] ET AL) 26 April 2007 (2007-04-26) paragraphs [0070] - [0079] ----- -/-	1-16



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

2 July 2012

Date of mailing of the international search report

09/07/2012

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Dupuis, Hervé

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2012/056303

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	BUNKE H ET AL: "Recent advances in graph-based pattern recognition with applications in document analysis", PATTERN RECOGNITION, ELSEVIER, GB, vol. 44, no. 5, 1 May 2011 (2011-05-01), pages 1057-1067, XP027595411, ISSN: 0031-3203, DOI: 10.1016/J.PATCOG.2010.11.015 [retrieved on 2011-01-11] page 1063, left-hand column -----	1-16
A	WO 2011/022660 A1 (PURETECH VENTURES LLC [US]; OLLE BERNAT [US]; BEHR JONATHAN ROBERT [US]) 24 February 2011 (2011-02-24) page 13, lines 16-35 -----	1-16
A	EP 1 587 012 A2 (MICROSOFT CORP [US]) 19 October 2005 (2005-10-19) paragraphs [0018], [0048] - [0053] -----	1-16

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2012/056303

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 03062943	A2	31-07-2003	AU 2003237982 A1 02-09-2003
		EP 1483725 A2 08-12-2004	
		US 2004204925 A1 14-10-2004	
		WO 03062943 A2 31-07-2003	

US 2007094066	A1	26-04-2007	US 2007094066 A1 26-04-2007
		US 2007094067 A1 26-04-2007	
		US 2007100680 A1 03-05-2007	
		US 2010324985 A1 23-12-2010	

WO 2011022660	A1	24-02-2011	US 2012149584 A1 14-06-2012
		WO 2011022660 A1 24-02-2011	

EP 1587012	A2	19-10-2005	CN 1694101 A 09-11-2005
		EP 1587012 A2 19-10-2005	
		JP 4838529 B2 14-12-2011	
		JP 2005302043 A 27-10-2005	
		KR 20060045782 A 17-05-2006	
		US 2005234972 A1 20-10-2005	
