Network restructuring after a node removal

Rokia Missaoui*

Université du Québec en Outaouais (UQO),
Case postale 1250, succursale Hull,
Gatineau (Québec), J8X 3X7, Canada
E-mail: rokia.missaoui@uqo.ca
*Corresponding author

Elsa Negre

Université Paris-Dauphine,
Place du Maréchal de Lattre de Tassigny,
75775 Paris Cedex 16, France
E-mail: elsa.negre@dauphine.fr

Dyah Anggraini

Université du Québec en Outaouais
E-mail: angd01@uqo.ca

Jean Vaillancourt

Université du Québec en Outaouais
E-mail: jean.vaillancourt@uqo.ca

Abstract: Central nodes (*i.e.*, prominent actors) in a social network are those that are linked to other nodes in an extensive or critical manner. Therefore, their removal may lead to points of failure.

The objective of the present work is to exploit network topology to devise an approach towards (i) finding substitute(s) to a deleted node, and (ii) adding appropriate links to avoid a fragmentation of the network into unconnected subgraphs. The approach exploits the role played by nodes to predict the new structure of a social network once one entity disappears. The role of a node in the network is expressed in terms of the number of interactions it has with the rest of the network. Three important roles are considered: the *leader*, the *mediator* and the *witness*. An entity acts as a leader, a mediator or a witness if it has a high degree centrality, betweenness centrality and closeness centrality, respectively.

Keywords: Social network analysis; Centrality; Node removal; Role; Network prediction

Reference to this paper should be made as follows: Missaoui, R. and Negre, E. and Anggraini, D. and Vaillancourt, J. (xxxx) 'Social Network Restructuring After a Node Removal', *International Journal of Web Engineering and Technology*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Rokia Missaoui received her Ph.D. in Computer Science (CS) in 1988 from Université de Montréal, Canada. She is a Full Professor in the Department of CS and Engineering, UQO, Canada. Before joining UQO, she was a professor at UQAM (Université du Québec à Montréal) between 1987 and 2002. Currently, she is the head of the LARIM laboratory. Her research interests include data mining and warehousing, formal concept analysis and social network analysis.

Elsa Negre received her Ph.D. in CS in 2009 from Université François-Rabelais de Tours, France and was a postdoctoral fellow at UQO in 2010-2011. She is currently an Assistant Professor at Université Paris-Dauphine, France. Her research interests include query recommendation and personalisation, data warehousing and social network analysis.

Dyah Anggraini is a Ph.D. student at UQO. Her research interests include social network analysis and data mining.

After earning his Ph.D. in Mathematics in 1987 from Carleton University, Jean Vaillancourt became a professor of Mathematics and Statistics at Université de Sherbrooke, where he subsequently held the position of Associate Dean of Science until 2001. He is currently President of the UQO in Gatineau and still pursues his research interests in stochastic modeling and data mining pertaining to various applications, notably social media. He is an associate member of the EMOSTA laboratory at UQAM in Montreal as well as a regular collaborator with the LARIM laboratory at UQO.

1 Introduction

Social network analysis (see (Carrington, 2007; Knoke and Yang, 2008; Wasserman and Faust, 1994)) is an important research area that has attracted many research communities and has been studied according to different approaches and techniques. A social network is a dynamic structure (generally represented as a graph) of a set of entities/actors (nodes) together with links (edges) between them.

Like all social structures, each actor plays a more or less important role within the network like the *leader* who interacts with many other entities or the *mediator* who acts as an intermediate entity between groups or the *witness* who has the best visibility about the information flow in the network.

Many studies in social network analysis (SNA) focus on static networks as indicated in Jamali and Abolhassani (2006). However, a social network is a dynamic structure where links and entities appear and disappear. In this paper we study the link prediction problem when a node is deleted. In practical terms, given a social network, what happens if an entity disappears? What will be the new structure of the network? Which node(s) will become substitute(s) and play the role of the deleted node if the latter happens to be a leader or a mediator or a witness? What are the links that more likely will be created in order to maintain the network in some stable structure?

The disappearance of *central* nodes (*i.e.*, the ones with at least one high centrality score) will very likely not have the same impact on the network as the disappearance of rather peripheral ones (*i.e.*, nodes with low centrality scores). Based on this fact, we propose a procedure for predicting the evolution of the structure of a social network after the deletion of an entity. Such a procedure first identifies the role of the removed node within the network, then effectively removes the node and its ties with other nodes, determines the potential substitute(s) and finally creates new links between the (unique) selected substitute (if it exists) and some identified nodes. Since degree, betweenness and closeness centralities are among the most important measures for the role or position of a node in a network, we will focus on them and associate to them three distinct roles, respectively: *Leader*, *Mediator* and *Witness.* When no substitute is found, the nodes formerly linked to the deleted node are connected to form a *clique.*

This paper is organized as follows: the next section gives some background on social network analysis and motivates our approach. Section 3 presents some related work while in Section 4, we describe our approach. An experimental study conducted on real large datasets is described in Section 5. Finally, Section 6 concludes this paper and provides future work.

2 Background

In the following we first recall some useful notions and then we illustrate through a simple example the way our approach works. The network is assumed to be a connected, undirected and unweighted graph.

Consider the $KITE$ network defined by David Krackhardt as shown in Figure 1 where ten individuals are linked through eighteen edges. One may first observe that nodes D, F, G and H are central nodes in that network.

2.1 Centrality measures

There are many measures of node centrality in a graph to capture the importance of an entity within such structure. For example, the *degree centrality*, the *betweenness centrality*, and the *closeness centrality* of a given node (see Wasserman and Faust (1994)) or a group of nodes (see Everett and Borgatti (2005)) are frequently used centrality measures.

Degree centrality for individual nodes provides the number of their direct links and helps identify *leaders* which have the (almost) highest number of links within the network. Group degree centrality represents the number of nodes outside the group that are linked to elements of the group. The *normalized* node degree centrality and group degree centrality in a given social network SN are computed as follows:
$C_D^{SN}(i) = \frac{d(i)}{n-1}$ for a node i
$C_D^{SN}(G) = \frac{|N(G)|}{n-|G|}$ for a group G of nodes,
where $d(i)$ is the degree (number of edges) of i, and $N(G)$ is the set of nodes

which do not belong to the group G but are adjacent to an element of the group.

Betweenness centrality expresses the amount of control that a node (or a group of vertices) possesses over the interactions of other nodes in the network. It is high for *mediators* (or *brokers*) which are nodes that act as intermediaries between other nodes or as joins between communities. The betweenness centrality indicator is computed as follows:

$C_B^{SN}(i) = \frac{2\times\sum_{i\neq j\neq k}\frac{p_{jk}(i)}{p_{jk}}}{(n-1)(n-2)}$ for a node i

$C_B^{SN}(G) = \frac{2\times\sum_{j<k}\frac{p_{jk}(G)}{p_{jk}}}{(n-|G|)(n-|G|-1)}$ for a group of nodes G

where p_{jk} is the number of shortest paths between nodes j and k and $p_{jk}(i)$ (resp. $p_{jk}(G)$) is the number of shortest paths between j and k crossing i (resp. G).

Closeness centrality of an individual node indicates how a node is close to the other nodes in the network and hence how fast information circulates from a given node to other reachable nodes in the network. This indicator will be used to identify what we call a *witness*. We define the role of witness in the network as being the one with a high closeness centrality, i.e., the one whose distance to other nodes is short within the network. Everett and Borgatti (2005) define group closeness centrality as "the normalized inverse sum of distances from the group to all node outside the group". The *normalized* closeness centrality and group closeness centrality are computed as follows:

$C_C^{SN}(i) = \frac{n-1}{\Sigma_{j=1}^{n} d(i,j) i\neq j}$ for a node i

$C_C^{SN}(G) = \frac{\Sigma_{i=1}^{n}[C_C^{SN}(N^*) - C_C^{SN}(i)]}{[(n-2)(n-1)]/(2n-3)}$ for a group of nodes G

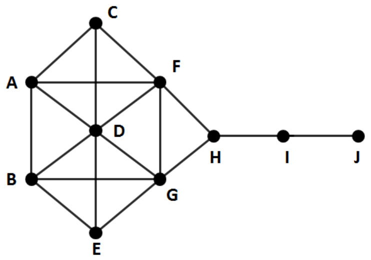
where $d(i,j)$ is the shortest path between i and j, and $C_C^{SN}(N^*)$ is the highest closeness centrality in the network.

Computing the betweenness and closeness centralities for all nodes in the network assumes the computation of the shortest paths between all pairs of nodes. For example, the calculation of betweenness centrality requires $O(n+m)$ space and runs in $O(nm)$ time for unweighted networks using the algorithm defined in Brandes (2001), where n and m are the number of nodes and links in the network respectively.

Figure 1 shows the value of degree, betweenness and closeness centralities for each node in the $KITE$ network.

2.2 Network measures

Network measures (Kaiser, 2008; Watts and Strogatz, 1998) are metrics that allow the quantification of a network as a whole and are helpful for comparing and classifying a set of graphs (e.g., samples of an initial graph) or when one network has undergone some topological changes (e.g., node/edge insertion or removal). The density, diameter, degree distribution and clustering coefficient are among the numerous network measures. The network diameter is the maximum length of shortest paths between two nodes while the density of a network is the proportion of the total number of links over the number of all possible links that can hold between existing nodes. The (node) degree distribution provides the number of



	Degree centrality	Betweenness centrality	Closeness centrality
A	0.444	0.023	0.52
B	0.444	0.023	0.52
C	0.333	0	0.5
D	**0.667**	0.102	0.6
E	0.333	0	0.5
F	0.556	0.231	**0.64**
G	0.556	0.231	**0.64**
H	0.333	**0.389**	0.6
I	0.222	0.222	0.42
J	0.111	0	0.31

Figure 1 **Left**: The *KITE* social network. **Right**: the centrality scores.

nodes with degree k for $k = 0, 1, \ldots$. The clustering coefficient of a network (see Watts and Strogatz (1998)) is the average value of the local clustering coefficients of all its n nodes.
$C_1 = \frac{1}{n}\sum_{i=1}^{n} C_i = \frac{1}{n}\sum_{i=1}^{n} \frac{2\Gamma_i}{k_i(k_i-1)}$.
where k_i is the number of the actual neighbors of node i while Γ_i represents the number of edges between the neighbors of node i. An adjusted (and more reliable) clustering coefficient is proposed in Kaiser (2008) where n in C_1 is replaced with the number of nodes that have at least two neighbors.

2.3 Motivation

As stated before, we propose a role-based approach with its associated procedure to predict the structure of a social network when a node disappears. Three typical cases of the role played by the deleted node are considered: *Leader*, *Mediator*, and *Witness*. If a node has none of these roles, then it is labeled *Other*. When an individual disappears from the network (e.g., account closure, person's retirement), all the links associated with such a node are removed. Then, one substitute of the deleted node will be sought, only when the vanished entity has played the role of a leader, mediator or witness in the network. The question is then the following: what is the impact of this disappearance on the remaining individuals? Will some individuals form a clique for example? This will depend on the structure of the network, the role previously played by the deleted node in that network as well as the desired linking option. We consider four possible options to link nodes in the network:

- *OLD*: a selected node i acting as a substitute is linked to the former neighbors of the deleted node X. E.g., when Bob replaces the group leader Martin, then he initiates a new tie with anyone (not already linked to Bob) who was linked to Martin.

- *IMP*: a selected node i is linked to a node z whenever the latter has a centrality (degree, betweenness or closeness) very close (*i.e.*, deviate by at most a ratio ε) to that of i. E.g., when Bob replaces the group leader Martin, then he initiates a new tie with anyone who has almost the same position as him in the organization, i.e., the one who has a degree centrality very close to his own centrality.

- *CLIQUE*: this option is only used when no substitute is found and aims to link the former neighbors of the deleted node X in a clique. E.g., when a central member of a family (or clan) such as a godfather or a patriarch passes away, the close collaborators or relatives strengthen their ties by forming a clique.
- *NGB*: a node i is linked to node z if the latter has a number of common neighbors with i very close to the maximum number of common neighbors that i shares with other nodes. E.g., if Jean and Daniel have many collaborators in common, then they will more likely collaborate (and link together) in the future.

Using the network shown in Figure 1, let us remove node B which turns to be *Other* because none of its three centrality scores deviates for example by a ratio $\varepsilon = 10\%$ from the corresponding highest value of degree, betweenness or closeness centrality. The graph becomes the one in the middle of Figure 2 after the deletion of B. In such a case, all the nodes that were linked to B will form a clique (the unique link option when the deleted node is *Other*) as indicated in Figure 2-c.

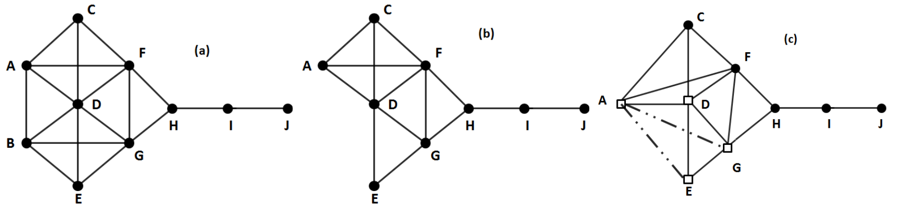


Figure 2 (a) The initial network. (b) AFTER deleting node B and related ties. (c) AFTER forming a clique with nodes A, D, E, and G that were neighbors of B.

If however node D is deleted, a substitute is sought because the role of D is a *Leader*. If we tolerate a deviation ratio $\varepsilon = 25\%$ from the degree centrality of D, then a node with a degree centrality at least equal to $0.667 \times 0.75 = 0.50$ is a potential substitute. This is the case of nodes F and G whose degree centrality becomes 0.50 in the new network deprived from D (see Table 1). If F is the selected substitute for D, then the new network will be Figure 3-(b) if the link option is *IMP* or Figure 3-(c) if the link option is *OLD*.

Another replacement alternative is to restrict the search for the substitute to the community of the deleted node because entities of the same community tend to share some common features. In our example, suppose that the set of nodes $\{A, B, C, D, E, F, G\}$ represents a community. Then, any search for the substitute of one of these nodes will be limited to its membership community. The benefit of such restriction lies in the fact that the exploration of the network and the computation of centrality measures are limited to the community to which the deleted node belongs to rather than to the entire network. However, there are some real-life situations where it is much more valuable to explore the whole network to better find a substitute with a role close to that of the deleted node.

Note that if the deleted node is a leaf (has only one link with the rest of the network), then there is no need to find a substitute or create new links between some existing nodes. In our example, the deletion of J from the initial network does not lead to the identification of a substitute or the insertion of new links.

3 Related work

Many social network evolution models have been proposed. However, they are mainly based on adding nodes or links. Dynamic networks are a kind of graphs (see (Ben-Naim and Krapivsky, 2007; Toivonen et al., 2009)) in which nodes and links can be added and/or deleted. In such networks, links are generally added/created by using the triadic closure and focal closure (see (Kumpula et al., 2007; Toivonen et al., 2009)) through non deterministic approaches generally limited to geodesic neighborhood. We recall that the triadic closure is the probability that B is linked to C knowing that A is already linked to B and C. The focal closure is the probability that A and B that share an interest could interact while the geodesic neighborhood of A is the set of nodes located at a geodesic distance (i.e., the shortest path) from A. The deletion of a node leads to the deletion of the adjacent links and the substitution of the deleted node by another one (see Toivonen et al. (2009)). However, when nodes are not much connected to the network, this substitution seems not necessary.

There are many social network prediction methods. The studies in Liben-Nowell and Kleinberg (2003) and Tylenda, Angelova and Bedathur (2009) focus on predicting new links but exclusively in static social networks. In Liben-Nowell and Kleinberg (2003), the objective is to define approaches to link prediction based on measures that exploit the proximity of nodes within a network. More precisely, the idea is to identify proximity measures that efficiently predict the new links that will likely happen in the future in a social network by assuming that two close nodes have a greater probability to be linked. An experimental comparison of a set of measures (e.g., shortest-path distance, Adamic-Adar distance and common neighbor measure) and their impact on the quality of link prediction is proposed. In Tylenda, Angelova and Bedathur (2009), the authors take into account the evolution of a network over time by integrating edge weights (potentially derived from temporal characteristics) into existing link prediction methods. They also investigate a new testing method to compute the performance of prediction algorithms in ranking the neighbor nodes of a selected node. Finally, the issue raised in Hussain and Ahmed (2008) is close to ours since the authors handle the problem of identifying the node of the network which will replace a given deleted node. To that end, they propose a Bayesian approach by first computing the posterior probability of each node in the network. The best substitute is then the one with the most closer posterior probability to that of the deleted node. Moreover, they take into account the hierarchical structure of nodes (a background knowledge not reflected in the network) so that the substitute needs to be of similar level of hierarchy (e.g. a Vice-president should be replaced by another Vice-president in the organization). However, the link prediction problem is not studied. In Negre, Missaoui and Vaillancourt (2011), a method to predict the new structure of a social network once a node is deleted is proposed. It consists to find one or possibly many substitute(s) for the deleted node based on the importance of the interactions that hold between nodes. New links can then be established in the network and any isolated node is discarded.

As stated by Callaway et al. (2000), an important feature of interconnections is their robustness - or fragility - to the deletion of network nodes. To study such property, the authors exploit percolation on graphs having general degree distribution and apply the theory of percolation to the study of network resilience. We plan to further explore this aspect to handle some critical issues in the future.

The present paper extends our previous work both in its structure and content: (i) algorithms are simplified and each involved function is detailed, (ii) the closeness centrality is introduced to define a new role called *witness*, and (iii) the quality of substitution in terms of “return to the normal” relies on a more elaborated set of network parameters such as density, number of links and edges, diameter and global clustering coefficient of the network.

4 Predicting a social network structure

The approach we describe in this paper is based on the role played by nodes in a social network in terms of their interactions with other nodes to predict the new structure of that network after the deletion of one of its vertices. We recall that this method relies on the following hypotheses: (i) the network is a connected, undirected and unweighted graph, and (ii) two entities of the network are compared on the basis of their centrality in the network.

Let assume that node X disappears from the network. Then, we would like to (i) identify the node(s) of the network that will replace X and play a similar role as X, and (ii) define the new structure of the network in terms of new and vanishing ties among nodes. Substitution is possible when X acts either as a leader, mediator or witness and when one node emerges as a substitute for the removed node.

4.1 Role

We define the role r_X^{SN} of a given node X within the network $SN = \langle V, E \rangle$ as an element of the finite set $R = \{\text{Leader, Mediator, Witness, Other}\}$ such that

$$r_X^{SN} = \begin{cases} \textit{Leader} & \text{if } \frac{C_D^R(X)}{maxL} \geq 1 - \varepsilon, \\ & \text{where } maxL \text{ is the maximal value of} \\ & \text{the observed degree centrality within} \\ & \text{SN and } \varepsilon \text{ is a tolerance ratio.} \\ \textit{Mediator} & \text{if } \frac{C_B^R(X)}{maxM} \geq 1 - \varepsilon, \\ & \text{where } maxM \text{ is the maximal value of} \\ & \text{the observed betweenness centrality} \\ \textit{Witness} & \text{if } \frac{C_C^R(X)}{maxC} \geq 1 - \varepsilon, \\ & \text{where } maxC \text{ is the maximal value of} \\ & \text{the observed closeness centrality} \\ \textit{Other} & \text{otherwise.} \end{cases}$$

The value of ε used throughout the paper is a user-defined ratio (e.g., 10%) that indicates to what extent a given value is allowed to at most deviate from another one. When for example ε is null for role assignment, this means that the role of leader (or mediator/witness) is assigned to the deleted node if it has the highest degree (or betweenness or closeness) centrality in the network. However, a non null value of ε allows (i.e., tolerates) a relative deviation of at most ε from the highest value of degree (or betweenness/closeness) centrality.

In case a node X has more than one role, it will take the role associated with its centrality that is most relatively close to the maximal corresponding centrality. For example, if node X has roles *leader* and *mediator*, it will be a *leader* if $\frac{C_D^{SN}(X)}{maxL} \geq \frac{C_B^{SN}(X)}{maxM}$. Otherwise, X will be a *mediator*.

From Figure 1, one can see that $maxL = 0.667$, $maxM = 0.389$ and $maxC = 0.64$. With a ratio $\varepsilon = 20\%$ of $maxL$, $maxM$ and $maxC$ respectively, the deleted node X has the role of a *leader* if it belongs to the set $\{D, F, G\}$, the role of a *mediator* if it corresponds to H, the role of a *witness* if it belongs to the set $\{A, B, D, F, G\}$, and *other* if it belongs to the set $\{C, E, I, J\}$. If the deleted node is F, it has the double role of *leader* and *witness*, and the computation of the two ratio $\frac{C_D^{SN}(X)}{maxL}$ and $\frac{C_C^{SN}(X)}{maxC}$ leads to $\frac{0.556}{0.667} < \frac{0.64}{0.64}$. Then, F will have the role of *witness*.

Deleting a node from the network leads to deleting associated links. Based on the existing studies on link prediction and social network evolution, we believe that predicting the social network structure after the deletion of a given node can not be limited to only the deletion (and possibly the insertion) of links but also to the identification of one or many nodes that can play a similar role as the deleted node. With this observation in mind and assuming that the network should return to a "normal state" after a node deletion, our approach proceeds in two steps: (i) predict, if possible, which node of the network will replace the deleted node X, and (ii) predict the new interactions that will appear within the network.
For instance, if a *mediator* is deleted, then a new candidate for mediation is sought. However, if the deleted node is *other*, then its role is not very important within the network and hence no substitute is required. However, new links will be added between nodes that were attached to the deleted.

As indicated earlier, our approach takes into account the topology of the network to predict its structure after the deletion of one node. The three following cases are then considered.

4.1.1 Leader

If the deleted node X is a *leader*, i.e., it has a high degree centrality, then its links to other nodes are also deleted, but the past interactions between nodes or groups of nodes via the deleted node should in some way persist through the newly selected *leader* using one of the previously described link options.

4.1.2 Mediator

If the deleted node X is a *mediator* between some nodes or groups of nodes, i.e., it has a high betweenness centrality, then its existing links with some other nodes are also deleted and a similar reasoning can then be conducted like in the first case.

4.1.3 Witness

If the deleted node X is a *witness* between some nodes or groups of nodes, i.e., it has a high closeness centrality, then its existing links with some other nodes are also deleted and a similar reasoning can then be conducted as in the first case.

When the deleted node is a *leader* or a *mediator* or a *witness* and there is a substitute, then the latter will be linked to some other nodes of the network according to one of three established link options (OLD, IMP, NGB) as defined by Procedure *LinkS* (see explanation below). However, if no substitute exists, then some nodes are linked with some other ones according to one of two possible link options ($CLIQUE$, IMP) using Procedure *LinkNoS* (see below). Then, the new network structure and the new central nodes are returned.

4.1.4 Other

If the deleted node has no role, i.e., it is neither a leader nor a mediator/witness, then its links are also deleted but no substitute is sought because the node is not enough important to be replaced by another node. However, the nodes that were previously attached to the deleted one will form a clique.
Once the new network structure is established, if some nodes are completely isolated, i.e., they are not linked to any other node, they are then deleted.

4.2 Algorithms

In this section, we propose a main procedure (see Algorithm 1) called *PredictStruct* for predicting the social network structure after the deletion of one of its nodes. The algorithm incorporates the two steps of our approach: the identification of the substitute of the deleted node, and then the link management.

Given a social network SN, a deleted node X, and a ratio ε, the main procedure returns both the new predicted network SN' (where X and its related links are removed while new links are added) and the new sets of leaders, mediators and witnesses of SN'. The ratio ε represents a tolerated relative deviation of a given value and is used as a parameter of some defined functions. For substitute identification (see *Substitute*$(X, SN, \varepsilon, Indic)$), ε shows how $Indic$ (e.g., degree centrality) value of a substitute could deviate from the indicator value of the deleted node. Using our illustrative example and $\varepsilon = 25\%$, the potential substitutes for the leader D are those having a degree centrality at least equal to $0.667 \times (1 - 0.25) = 0.50$ in the network deprived from the deleted node and its associated links. This is the case of nodes F and G (see Table 1). The parameter ε used in functions *LinkS* and *LinkNoS* represents the deviation of a measure (e.g., the number of common neighbors) that a node w is allowed to have in order to be linked to i. It is ignored whenever the link option is CLIQUE.
After storing the number of neighbors NB of node X, the next executed instruction in the main procedure identifies the role of X (as defined in Subsection 4.1) via the function *Role*(X, SN, ε) which returns either *Leader*, *Mediator*, *Witness* or *Other* (Line 4). When X is a *leader* (*mediator* or *witness*), the parameter $Indic$ takes the value 'D' ('B' or 'C' resp.) that stands for degree (betweenness or closeness resp.) centrality (Line 5). For each one of these three main roles (leader, mediator and witness), the procedure looks for a substitute to

put in the set NR via *Substitute*$(X, SN, \varepsilon, Indic)$ (Line 7 and Algorithm 2). If NR is not empty, this means that a substitute exists and hence will be linked to some other nodes of the network via the function *LinkS*$(NR, SN', OpLink, \varepsilon, NB)$ where NB is the set of nodes previously linked to the deleted node X (see Line 9 and further explanations). Otherwise (i.e., no substitute is found), the procedure links nodes of the network via the function *LinkNoS*$(SN', OpLink, \varepsilon, NB)$ (Line 11). When the deleted node X has none of the three main roles, the nodes that were attached to it (i.e., those in the set NB) will form a clique as indicated in the *otherwise* part of the *switch* block (Line 5). Any isolated node is deleted (see Lines 14-15). Finally, the procedure computes the value of the three indicators for nodes of the predicted network SN' and returns SN' as well as the central nodes of the new network SN' (Lines 17-19).

```
 1: Procedure PredictStruct
 2:    Input  : $SN = \langle V, E \rangle$: initial network, $X$: deleted node, $\varepsilon$: tolerance ratio,
    $OpLink$: parameter for possible link options.
    Output: $SN'$ (a global variable): new network; $L, M$, and $W$: sets of leaders,
    mediators
             and Witnesses in $SN'$.
 3: $NB \leftarrow Neighbor(X)$; $L, M, W \leftarrow \emptyset$
 4: $R \leftarrow Role(X, SN, \varepsilon)$
 5: switch the role of $X$ do
    case $R =$ Leader $Indic \leftarrow$ ‘$D'$;
    case $R =$ Mediator $Indic \leftarrow$ ‘$B'$;
    case $R =$ Witness $Indic \leftarrow$ ‘$C'$;
    otherwise
       $SN' \leftarrow \langle V' = V - \{X\}, E' = E - \{e_{Xj}, \forall j \in V\} \rangle \cup$
       $LinkNoS(SN, “CLIQUE'', \varepsilon, X)$
    endsw
    endsw
 6: if $R \neq$ Other then
 7:    $NR \leftarrow Substitute(X, SN', \varepsilon, Indic)$
 8:    if $NR \neq \emptyset$ then
 9:       $SN' \leftarrow SN' \cup LinkS(NR, SN', OpLink, \varepsilon, NB)$
10:    else
11:       $SN' \leftarrow LinkNoS(SN', OpLink, \varepsilon, NB)$
12:    end if
13: end if
14: if $\exists y \in V' | \forall z \in V', z \neq y, \nexists e_{yz} \in E'$ then
15:    $V' \leftarrow V' - \{y\}$ //delete isolated nodes
16: end if
17: $L \leftarrow Leaders(SN')$
18: $M \leftarrow Mediators(SN')$
19: $W \leftarrow Witnesses(SN')$
20: return $(SN', L, M, W)$
```

Algorithm 1: Predicting the structure of a network.

Algorithm 2 describes the procedure *Substitute*$(X, SN, \varepsilon, Indic)$. Given a social network SN, a deleted node X, a threshold ε and an indicator $Indic$, the procedure returns a unique node as a substitute for X. For that purpose, the procedure computes the value of the indicator $Indic$ of the nodes in the network $SN' = \langle V', E' \rangle$, i.e., without X and its associated links (Lines 4-5) and stores in NR the potential substitutes having an indicator value in SN' close to (i.e., deviating at most by a relative proportion ε from) the indicator value of X in the initial network SN (Lines 6-11). If NR contains more than one substitute, only one node z with the maximal indicator value is returned as a substitute of X (Lines 12-14). The set NR containing the selected substitute for X is finally returned (Line 15).

```
1:  **Procedure** SUBSTITUTE
2:     **Input**  : $SN = \langle V, E \rangle$: initial network, $X$: deleted node, $\varepsilon$: tolerance ratio,
    $Indic$: degree or betweenness or closeness centrality indicator.
    **Output**: $NR$: set of potential substitutes for $X$.
3:  $NR \leftarrow \emptyset$
4:  $ValX \leftarrow ValIndic(X, SN, Indic)$
5:  $SN' \leftarrow \langle V' = V - \{X\}, E' = E - \{e_{Xj}, \forall j \in V\} \rangle$
6:  **for** each $w \in V'$ **do**
7:     $ValW \leftarrow ValIndic(w, SN', Indic)$
8:     **if** $ValW \geq (1 - \varepsilon) \times ValX$ **then**
9:        $NR \leftarrow NR \cup \{w\}$
10:    **end if**
11: **end for**
12: **if** $|NR| > 1$ **then**
13:    $NR \leftarrow Max(ValIndic(z, SN', Indic)$ )
14: **end if**
15: **return** $NR$
```

Algorithm 2: Finding the substitute of a deleted node.

4.3 Additional Functions

Functions *Role*, *ValIndic*, *LinkS* and *LinkNoS* can be briefly described as follows.

Role(i, SN, ε) returns the role of a given node i within the network SN by looking for the maximal values $maxL$, $maxM$ and $maxC$ among the network nodes and by comparing them to the threshold ε (as defined in Subsection 4.1).

$ValIndic(i, SN, Indic)$ returns the value of the indicator $Indic$ (the degree/betweenness/closeness centrality) of the node i within the network SN.

$LinkS(i, SN', OpLink, \varepsilon, NB)$ links the substitute node i in SN to other nodes in different manners depending on the chosen option $OpLink$. The considered options are: IMP, OLD and NGB. IMP links i to a node z when $C^{SN'}_{Indic}(z)$ deviates at most by a proportion ε from $C^{SN'}_{Indic}(i)$ in SN'. OLD links i to nodes in NB (neighbors of X) in order to maintain the previous interactions. NGB links i to node z if the latter has a number of common neighbors with i that deviates by

	Degree centrality	Betweenness centrality	Closeness centrality
A	0.375	0.071	0.5
B	0.375	0.071	0.5
C	0.250	0	0.44
E	0.250	0	0.44
F	0.5	0.286	0.61
G	0.5	0.286	0.61
H	0.375	0.429	0.61
I	0.250	0.250	0.44
J	0.125	0	0.32

(a) After the deletion of D

	Degree centrality	Betweenness centrality	Closeness centrality
A	0.375	0.029	0.50
B	0.50	0.095	0.53
C	0.25	0	0.44
E	0.25	0	0.44
F	0.625	0.36	0.666
G	0.50	0.19	0.61
H	0.375	0.428	0.615
I	0.250	0.250	0.44
J	0.125	0	0.32

(b) After using IMP with $\varepsilon = 25\%$.

Table 1 Indicator values for SN' after deleting D and using IMP to create new links.

at most a ratio ε of p, where p is the maximum number of common neighbors that i shares with other nodes in the new network SN'.

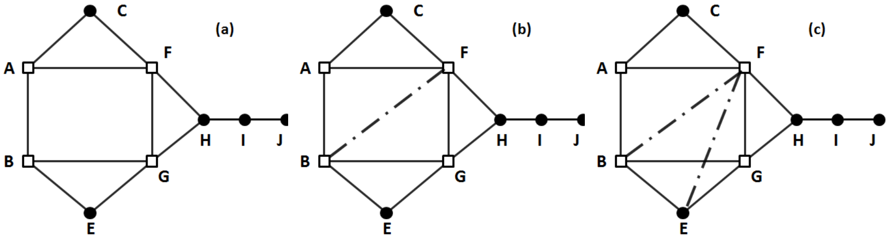


Figure 3 (a) The new network AFTER deleting node D and related ties. (b) AFTER using the link option IMP with $\varepsilon = 25\%$. (c) AFTER using the link option OLD.

LinkNoS(SN', OpLink, ε, NB) is used when no substitute is found and aims to create links between nodes in NB (neighbors of node X). The kind of the link depends on the selected option of *OpLink* which can be either $CLIQUE$ or IMP. The option $CLIQUE$ forms a clique with the identified nodes while the option IMP has the same meaning as IMP used in *LinkS*.

We are aware that the proposed approach handles typical situations rather than every possible situation. However, our approach can work decently in some extreme situations like in *complete networks* where every node is linked to all the remaining nodes, or in *star graphs* in which only one node is linked to the rest of the nodes. For complete graphs with n nodes, the algorithm returns a new complete graph with $(n-1)$ nodes, deprived of the deleted node and its associated links. For star graphs with n nodes, no substitute of the central node exists. In such a case, the unique link option consists in forming a clique with the remaining nodes.

As an illustration, let us consider the network SN given in Figure 1 and let apply Procedure $PredictStruct(SN,\ X,\ \varepsilon,\ OpLink)$ with the following values: $X = D$, $\varepsilon = 0.25$, and $OpLink = OLD$. The node D to delete has the role of a leader since $0.667 - C_D^{SN}(D) = 0$ which is less than 0.25. The variable $Indic$ will then take the value 'D' (degree centrality). When Procedure $Substitute(D, SN', 0.25, \text{'D'})$ is called, node D and its associated links are first deleted from SN as depicted in Figure 3-(a). Then, nodes F and G are potential substitutes because their degree centrality in the new network SN' (see the new scores in Table 1) is equal to 0.50 which is equal to $0.667 \times (1 - 0.25)$. Let assume that F is selected as the substitute for D. When Procedure $LinkS(NR, SN', OpLink, \varepsilon, NB)$ is called with the following values: $NR = \{F\}$, $OpLink = IMP$, $\varepsilon = 0.25$ and $NB = \{A, B, C, E, F, G\}$, we get the network SN' displayed in Figure 3-(b) where F is newly linked to node B which is the only node in NB that has a degree centrality at least equal to $(1 - 0.25) \times 0.50$. If however $OpLink = OLD$, F is automatically linked to the nodes in NB that are not already linked to it (see Figure 3-(c)).

4.4 Community detection

The detection of communities in a network is an important issue in social network analysis (see Fortunato (2010) for a survey) and has attracted many researchers in sociology, biology, computer science, and so on. A community is a kind of cluster where many edges link nodes of the same cluster and few edges link nodes of different clusters. A commonly used approach to find communities is based on betweenness centrality (see Girvan and Newman (2002)) which avoids having isolated nodes but has high computational requirements. Another commonly used method is based on the modularity maximization (see Newman (2004)) which calculates the quality of a particular clustering of a network into communities. Some further optimizations have been proposed. One of them is a parameter-free and easy-to-use approach described in Chen, Zaïane and Goebel (2009). Recent studies focus also on community evolution. For instance, Lin et al. (2009) take into account the known communities at time t to determine the communities at time $t + 1$.

Looking for the substitutes within the whole network can be tedious and expensive. Let us consider for example a company where one of the leaders leaves (e.g., retirement or firing). Then, the company will look for a substitute that has many interactions with individuals either within the community of the individual that left the company, or even outside his own community. Restricting the search for the substitutes in a limited part of the network will reduce the processing time but assumes that a preprocessing of the network is conducted for community detection.

A possible improvement of our approach is then to first determine the *cluster* or the *block* (of equivalent elements) in which the deleted node X holds to further restrict the search of substitutes to such group. Finding blocks of structurally equivalent elements from the network is done through the process of blockmodeling as described in White, Boorman and Breiger (1976), which also allows the construction of a smaller comprehensible structure.

5 Experiments

In this section, we empirically evaluate the potential of our approach for predicting the structure of the network when a node disappears. The experiments are conducted with three specific objectives in mind: (i) performance evaluation, (ii) substitution evaluation, and (iii) network perturbation. To that end we use two real datasets detailed further. Our prototype is implemented in Java and the tests were conducted on a Core 2 Duo E6750 with 2.66GHz and 3.23Go of RAM running under Windows XP.

We use two commonly known and large undirected and unweighted networks that we respectively call $COAUTHOR$ and $POWER\ GRID$ available at http://www-personal.umich.edu/ mejn/netdata/.

The dataset $COAUTHOR$ is a co-authorship network of scientists working on network theory and experiment, as established by Newman (see Newman (2006)). It contains 1589 nodes and 2742 links. The dataset $POWER\ GRID$ represents the topology of the Western States Power Grid of the United States. The 4941 entities are transformers, substations, and so on while the 6594 interactions are high-voltage transmission lines. Hence, $POWER\ GRID$ is about three times larger than $COAUTHOR$.

5.1 Performance Analysis

The experiments conducted on the two datasets and presented here allow to evaluate the time needed to predict the network structure after the deletion of one of its nodes. Note that these two datasets have been evaluated for an important set of nodes, for the two roles *Leader*, and *Mediator*, and for *Other*, for each possible option when substitutes exist ("WITH substitutes" on Figure 4) or not ("WITHOUT substitutes"). The overall mean time for the link options and the three role cases is given through the "WITH and WITHOUT" chart on Figure 4. One may notice that this average time is slightly smaller than the case of "WITHOUT substitutes" due to the fact that the overall cost is biased by the small processing time of nodes in case of *Other*. Moreover, knowing that the three roles (*Leader*, *Mediator* and *Witness*) are similarly processed within the algorithm, performance results are similar for these cases. Thus, no differentiation between these roles is provided in the empirical results. Finally, note that the execution time here is the CPU time needed for detecting substitutes and adding new links. The computation of indicator values for each node is done outside our prototype using $UCINET$ (available via http://www.analytictech.com/ucinet/).

As expected, Figure 4 shows that the execution time to predict the network structure after the deletion of one of its nodes increases with the network size. Moreover, this time is more important when no substitute exists than when a substitute or a group of substitutes is found. This is due to the fact that in the former case, all the network will be explored while in the latter case, a subset of the network is generally considered (mainly substitutes and their neighbors and

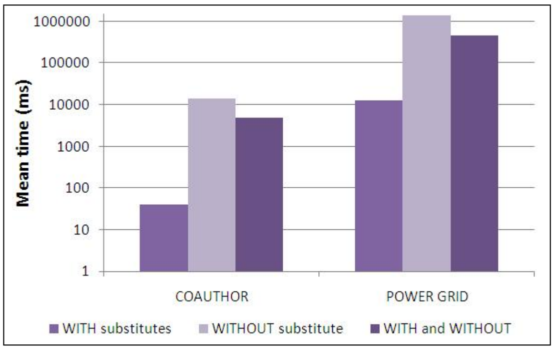


Figure 4 Mean time (in milliseconds) when substitutes exist or not.

sometimes the nodes linked to the removed node when OLD option is retained). Note also that the high execution times for the $POWER\,GRID$ dataset in Figure 4 are mainly due to the option NGB which computes the common neighbors of two nodes in the whole network. However, the merit of the option NGB is its ability to provide a good prediction based on the common neighbor measure as stated in Liben-Nowell and Kleinberg (2003).

5.2 Substitute Quality Analysis

To evaluate the quality of the prediction of a substitute and mainly the impact of ε on the identification of substitutes, we use the classical recall R and precision P measures as well as the F-measure F which are defined as follows:

$$R = \frac{|S_{rel} \cap S_{retri}|}{|S_{rel}|}\ ,\ P = \frac{|S_{rel} \cap S_{retri}|}{|S_{retri}|}\ \ F{=}\frac{2\times(R\times precision)}{(R+P)}$$

where S_{rel} is the set of relevant substitutes (i.e., those that occur when ε is almost null) and S_{retri} is the set of substitutes retrieved by our main procedure with a tolerance ratio $0 \leq \varepsilon \leq 1$.

Figure 5 displays the recorded F-measure for the two datasets according to the value of ε. It shows that the curves have a classical appearance and that smaller the ratio ε is (i.e., the higher is $1-\varepsilon$), the better is the F-measure. For the two datasets, the F-measure is higher than 0.9 for $1-\varepsilon \geq 75\%$, which is an indication that ε can reach up to 25% and still lead to a good substitute prediction. The absence of values for $COAUTHOR$ between 100% and 80% is due to the absence of substitutes.

5.3 Network Perturbation Analysis

The most popular (and easy to interpret) network measures for comparing two given graphs are single values (e.g., clustering coefficient) rather than distributions (e.g., the degree distribution) as stated by Kaiser (2008). To analyze the perturbation of the network following a node removal and the conducted postprocessing (substitute identification and link addition), we look at the mean variation (gain or loss) of five network measures before and after node deletion: density, diameter, global clustering coefficient as well as the number of nodes and

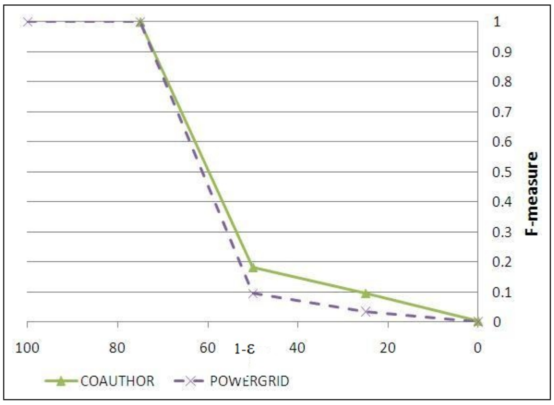


Figure 5 F-measure of the substitutes for the two datasets according to ε

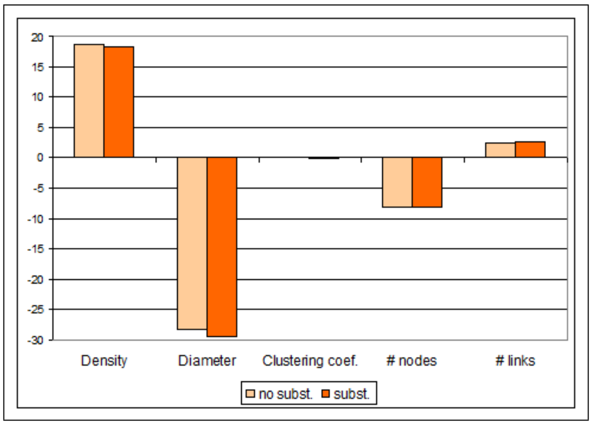


Figure 6 Network perturbation: $COAUTHOR$, $\varepsilon = 20\%$, and with or without substitutes.

links. In fact, a variation close to 0 is sought because it means an almost null perturbation and hence "a return to normal" of the network following a node deletion.

Figures 6 and 7 exhibit the gain/loss in network measures following a node removal for the dataset COAUTHOR with $\varepsilon = 20\%$ and some variants. Figure 6 shows that the charts look similar in the two cases: presence vs absence of substitutes. In both cases, the clustering coefficient has almost no variation. However, the density (resp. diameter) of the network has a relatively important increase of 18% (decrease of 27%) following node removal. Figure 7 shows the impact of the link options (IMP, NGB, OLD and CLIQUE) on the network perturbation. The best results (i.e., no variation) are obtained for OLD and IMP. However, NGB and $CLIQUE$ options seem to be the worst ones because the network variations are the most important for the five measures.

We believe that additional and more extensive experiments are needed to validate these preliminary results under various situations.

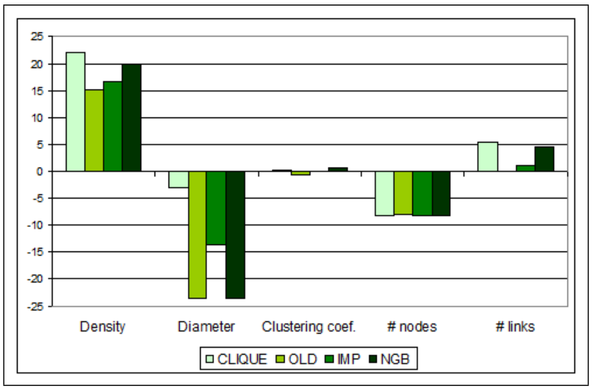


Figure 7 Network perturbation: *COAUTHOR*, $\varepsilon = 20\%$, and link options.

To the best of our knowledge, there is no study that handles the same issue as ours, except the work in Hussain and Ahmed (2008) that seeks for a substitute of a deleted node without considering the new links to create. Therefore, no comparative study between our method and other existing methods could be conducted.

6 Conclusion and future work

In this paper, we have proposed a non-probabilistic approach for social network structure prediction once a node is deleted from the network. This approach, inspired from human behavior in social and professional situations, is a first step toward a prediction method of social network structure when one entity disappears. It is based on the role (*Leader*, *Mediator* or *Witness*) played by entities in terms of the ties they maintain in the network.

The preliminary experiments conducted on two known social networks show that our system can predict a social network structure in a reasonable time and that the options OLD and IMP offer relatively good results in terms of execution time and precision (quality of prediction). However, option NGB is the worst both in terms of execution time and precision, mainly when no substitute is found.

Our future work includes the following issues: (i) study and quantify the degree of influence/prestige of a node in directed networks rather than the role (based on centralities) in undirected graphs, (ii) devise a procedure to predict a potential deletion propagation (*i.e.*, cascading effect) to other nodes (e.g., when an influential entity leaves an organization), and (iii) handle the removal of a set of nodes rather than only one. Finally, we plan to explore the potential of statistical distributions to handle network structure prediction.

Acknowledgment

We acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

This paper is a revised and expanded version of a paper entitled *Predicting a social network structure once a node is deleted* presented at ASONAM'2011, Taiwan, July 2011.

References

Ben-Naim, E. and Krapivsky, P.L. (2007) 'Addition - deletion networks', *Journal of Physics A: Mathematical and Theoretical*, 40(30):8607.

Brandes, U. (2001) 'A faster algorithm for betweenness centrality', *Journal of Mathematical Sociology*, 25:163–177.

Callaway, D.S., Newman, M.E.J., Strogatz, S.H. and Watts, D.J. (2000) 'Network Robustness and Fragility: Percolation on Random Graphs', *Physical Review Letters*, 85:5468–5471.

Carrington, P.J., Scott, J. and Wasserman, S. (Eds.), (2005) *Models and methods in social network analysis*, Cambridge Univ. Press, Cambridge.

Chen, J., Zaïane, O.R. and Goebel, R. (2009) 'Detecting communities in social networks using max-min modularity', *SDM*, pages 978–989.

Everett, M.G. and Borgatti, S.P. (2005) 'Models and methods in social network analysis', in Carrington, P.J., Scott, J. and Wasserman, S. (Eds.), *Models and methods in social network analysis*, Cambridge Univ. Press, Cambridge, pages 57–76.

Fortunato, S. (2010) *Community detection in graphs*, Physics Reports, 486(3-5):75–174.

Girvan, M. and Newman, M.E.J. (2002) 'Community structure in social and biological networks', in *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826.

Hussain, D.M.A. and Ahmed, Z. (2008) 'Dynamical adaptation in terrorist cells/networks', in *SCSS (2)*, pages 557–562.

Jamali, M. and Abolhassani, H. (2006) 'Different aspects of social network analysis', in *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 66–72.

Kaiser, M. (2008) 'Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures for small-world networks', *New Journal of Physics*, Vol. 10, No. 8, 1–11.

Knoke, D. and Yang, S. (2008) *Social Network Analysis*, Sage, second edition.

Kumpula, J.M., Onnela, J.P., Saramäki, J., Kaski, K. and Kertész, J. (2007) 'Emergence of communities in weighted networks', *Physical Review Letters*, 99(22).

Liben-Nowell, D. and Kleinberg, J. (2003) 'The link prediction problem for social networks', in *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, NY, USA, pages 556–559.

Lin, Y.R., Chi, Y., Zhu, S., Sundaram, H. and Tseng, B.L. (2009) 'Analyzing communities and their evolutions in dynamic social networks', in *ACM Trans. Knowl. Discov. Data*, 3:8:1–8:31.

Negre, E., Missaoui, R. and Vaillancourt, J. (2011) 'Predicting a social network structure once a node is deleted', in *ASONAM*, pages 297–304.

Newman, M.E.J. (2004) 'Fast algorithm for detecting community structure in networks', *Phys. Rev. E*, 69(6):066133.

Newman, M.E.J. (2006) 'Coauthorship networks and patterns of scientific collaboration', *Physical Review E*, 74:036104.

Toivonen, R., Kovanen, L., Kivel, M., Onnela, J.P., Saramki, J. and Kaski, K. (2009) 'A comparative study of social network models: Network evolution models and nodal attribute models', *Social Networks*, 31(4):240–254.

Tylenda, T., Angelova, R. and Bedathur, S. (2009) 'Towards time-aware link prediction in evolving social networks', in *The 3rd SNA-KDD Workshop '09 (SNA-KDD'09)*, SIGKDD.

Wasserman, S. and Faust, K. (1994) *Social network analysis : methods and applications*, Cambridge University Press, 1st edition.

Watts, D.J. and Strogatz, S.H. (1998) 'Collective dynamics of small-world networks', *Nature*, 393:440–442.

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

Zhu, H. and Zhou, M. (2009) 'M-m role-transfer problems and their solutions', *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 39(2):448–459.