Predicting a Social Network Structure Once a Node is Deleted

Elsa Negre
Université du Québec en Outaouais
Case postale 1250, succursale Hull
Gatineau (Québec), J8X 3X7
Canada
Email: Elsa.Negre@uqo.ca

Rokia Missaoui
Université du Québec en Outaouais
Case postale 1250, succursale Hull
Gatineau (Québec), J8X 3X7
Canada
Email: Rokia.Missaoui@uqo.ca

Jean Vaillancourt
Université du Québec en Outaouais
Case postale 1250, succursale Hull
Gatineau (Québec), J8X 3X7
Canada
Email: Jean.Vaillancourt@uqo.ca

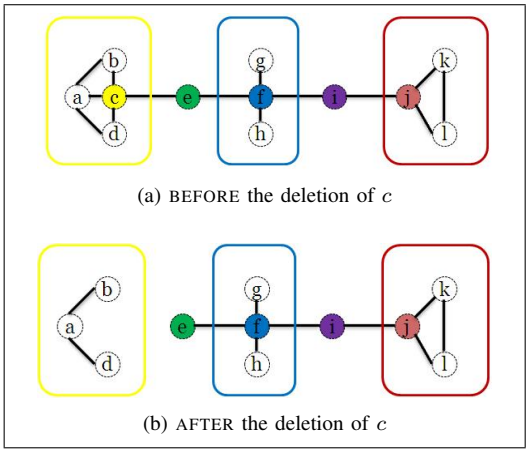
***Abstract*—Social networks are dynamic structures in which entities and links appear and disappear for different reasons. Starting from the observation that each entity has a more or less important role within the network, the objective of this article is to propose a method which exploits the role 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. Two roles are considered: the leader and the mediator with their corresponding measure: the degree centrality and the betweenness centrality.**

I. Introduction

Social network analysis [1] is an important research area that attracts many research communities and is being handled according to different approaches and techniques. A social network is a dynamic structure (generally represented as a graph) of a set of entities (nodes) together with links (edges) between them. Like all social structures, each entity plays a more or less important role within the network like the *leader* which interacts with many other entities or the *mediator* which acts as an intermediate entity between groups.

Most of studies on social network analysis (SNA) focus on static networks [2]. 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 nodes will play the role of the deleted node if the latter happens to be a leader or a mediator? What are the links that more likely will be created or deleted? Will some entities disappear from the network?

Since the disappearance of a *leader* will not have the same impact on the network as the disappearance of a rather isolated entity, we propose a method for predicting the evolution of the structure of a social network after the deletion of an entity by using an approach based on the role played by entities within the network. This approach relies on the following hypotheses: (i) the network is connected, (ii) the network is an undirected and unweighted graph, and (iii) two entities of the network are compared on the basis of the interactions they maintain with the rest of the network.



(a) BEFORE the deletion of c

(b) AFTER the deletion of c

Fig. 1. Undirected graph representing RS

Note that we particularly focus on entities having a special role (*leader* or *mediator*) and that our choices are inspired by real-life applications.

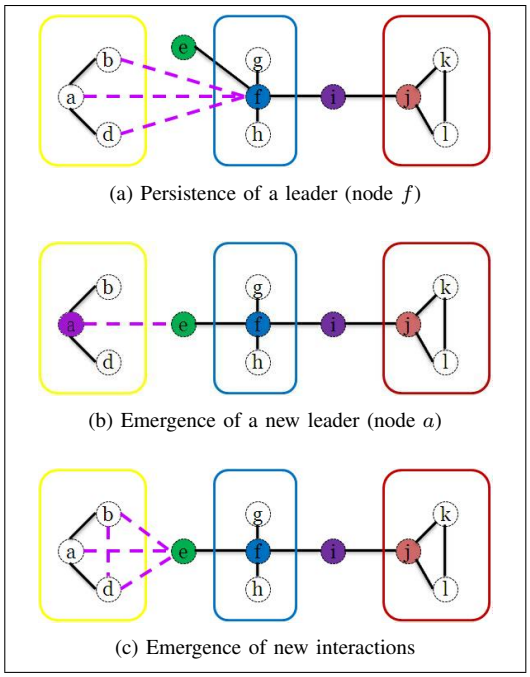
This paper is organized as follows: the next section motivates our approach through a toy example. Section III presents some related work while in Section IV, we describe our approach. An experimental study conducted on real datasets is described in Section V. Finally, Section VI concludes this paper and provides future work.

II. Motivation and Intuition

In this section, we illustrate through a toy example the way our approach works. We assume that a deleted entity X will be replaced by another entity (or even a set of entities) having a similar behavior as X in terms of interactions with the rest of the network.

Consider a simple social network $RS = \langle E, V \rangle$ (see Figure 1-a) of twelve individuals in $V = \{a, b, c, d, e, f, g, h, i, j, k, l\}$ who interact together as described in $E = \{(a,b), (a,c), (a,d), (b,c), (c,d), (c,e), (e,f), (f,g), (f,h), (f,i), (i,j), (j,k), (j,l), (k,l)\}$ where an edge (u, v) indicates that the individuals u and v are linked.

Now let assume that an individual disappears from the network (e.g., account closure, person's death or retirement or firing). First, the corresponding node is deleted from the



(a) Persistence of a leader (node f)

(b) Emergence of a new leader (node a)

(c) Emergence of new interactions

Fig. 2. Three predictions of network structure after the deletion of node c.

network as well as all the interactions associated with such a node. Then, one or many substitutes of the deleted node will be sought, mainly in case the vanished entity has played the role of a leader or a mediator in the network. The question is 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 and on the role of the deleted node in that network.

Assume the disappearance of the node c (see Figure 1-a), which is, in our example, a *leader*, i.e., a node having the most important number of interactions with the other nodes of the network. The graph becomes the one in Figure 1-b after the deletion of c. A new *leader* can then emerge from existing nodes and may inherit links that the deleted node had with other nodes. In our example, after the deletion of c, the node f is the one having the higher number of interactions with the remaining nodes. Node f can then be a substitute for c and the former interactions between c and the other nodes can be transferred to f (see Figure 2-a).

Another replacement alternative is to restrict the search for the substitute(s) to the community of the deleted node because entities of the same community tend to share some common interests. In our example, the community of c contains the nodes a, b and d. After the deletion of c, the node a is the one having the highest number of interactions with the remaining nodes. Node a can then replace the former leader c and the existing interactions between c and the other nodes are added to a (see Figure 2-b).

Another possibility is the occurrence of new interactions without the emergence of a *leader*. It could be the case of a friend group where, after the departure of a leader, the remaining friends mutually enforce their interactions. In our example, c was linked to a, b and d. After the deletion of c, the three remaining nodes are going to interact unless one of them becomes a *leader* (see Figure 2-c).

Notice that if the deleted node has a weak interaction with the rest of the network and therefore is neither a leader nor a mediator, then there is no need to find a substitute or create new links between some existing nodes. In our example, the deletion of h from the initial network does not lead to the identification of a substitute or the insertion of new links.

In this paper, we propose a role-based method and procedures 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 other. The last one occurs when a node is neither a leader nor a mediator.

III. 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 [3], [4] in which nodes and links can be added and/or deleted. In such networks, links are generally added/created by combining triadic closure and focal closure [4] through non deterministic approaches generally limited to geodesic neighborhood[1]. The deletion of a node leads to the deletion of the adjacent links and the substitution of the deleted node by another one [4]. 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 [5] and [6] focus on predicting new links but exclusively in non-dynamic social networks. In [5], 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 [6], 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 [7] is close to ours since the authors propose a Bayesian approach to predict which node of the network will replace a given deleted node. However, the link prediction problem is not studied and the authors focus on networks where nodes are organized into a hierarchy so that the "substitute" and the "deleted" node are preferably at the same level of the hierarchy.

In this paper, we propose a method to predict the new structure of a social network once a node is deleted. It consists

[1]Note 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, interact. The geodesic neighborhood of A is the set of nodes located at a geodesic distance from A (in terms of link number).

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.

IV. Predicting a social network structure

In this paper, we propose a non-probabilistic approach 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 proposed solution needs to keep the network connected, (ii) the network is an undirected and unweighted graph, and (iii) two entities of the network are compared on the basis of the interactions they maintain with the rest of the network.

Our problem can be stated as follows: assume that node X disappears from the network, we would like to (i) identify the node(s) of the network that will replace X and (ii) define the new structure of the network in terms of new and vanishing interactions among nodes. Substitution is possible when X acts either as a leader or a mediator and when one or many nodes emerge as substitute(s).

A. Definitions

In the following we first recall definitions and formulae known in the literature to further present our approach and related notions.

1) Indicators: Many indicators were proposed in the literature in order to capture the importance of an entity within the network. For example, the *degree centrality* and the *betweenness centrality* of a given node [8] or a group of nodes [9] are valuable measures to use in a network $RS = \langle E, V \rangle$ of n nodes.

Degree centrality for individual nodes helps identify central nodes or *leaders* which have the 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 [9]. The *normalized* degree centrality and group degree centrality are computed as follows:

$C_D^{RS}(i) = \frac{d(i)}{n-1}$ for a node i

$C_D^{RS}(G) = \frac{|N(G)|}{n-|G|}$ 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 is useful to identify *mediators* which are nodes that act as intermediate entities between other nodes. The betweenness centrality indicator[2] is computed as follows:

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

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

where p_{jk} is the number of shortest paths between 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).

[2]An effective fast algorithm has been proposed by [10] to compute the betweenness centrality. Its complexity is $O(nm + n^2 \log n)$ where m and n are the number of links and nodes in the network.

	Degree centrality	Betweenness centrality
a	0.273	0.09
b	0.182	0
c	**0.364**	0.445
d	0.182	0
e	0.182	0.509
f	**0.364**	**0.709**
g	0.091	0
h	0.091	0
i	0.182	0.436
j	0.273	0.327
k	0.182	0
l	0.182	0

TABLE I
Degree and betweenness centrality values of the network RS.

Using our illustrative example presented in Section II, Table I shows the value of degree centrality and betweenness centrality for each node in the network.

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

$$r_X^{RS} = \begin{cases} \text{Leader} & \text{if } maxL - C_D^{RS}(X) \leq limL, \\ & \text{where } maxL \text{ is the maximal value of} \\ & \text{the observed degree centrality within} \\ & \text{RS and } limL \text{ is a given threshold.} \\ \text{Mediator} & \text{if } maxM - C_B^{RS}(X) \leq limM, \\ & \text{where } maxM \text{ is the maximal value of} \\ & \text{the observed betweenness centrality} \\ & \text{and } limM \text{ is a given threshold.} \\ \text{Other} & \text{otherwise.} \end{cases}$$

The value of $limL$ (resp. $limM$) helps consider a deleted node as a leader (resp. mediator) whenever its degree centrality (resp. betweenness centrality) is in the interval $[maxL$ - $limL$, $maxL]$ (resp. $[maxM$ - $limM$, $maxM]$). The threshold $limL$ (resp. $limM$) can be perceived as the maximal deviation allowed from $maxL$ (resp. $maxM$) and can be determined by computing a user-defined ratio p of the maximal value $maxL$ (resp. $maxM$), i.e., $limL = p \times maxL$ (resp. $limM = p \times maxM$). When for example $limL$ is null, this means that the role of leader is assigned to the deleted node if it has the highest degree centrality in the network. However, a non null value of $limL$ allows (i.e., tolerates) a deviation from the highest value of degree centrality.

From Table I, one can see that $maxL = 0.364$ and $maxM = 0.709$. With a ratio p equal to 30% of $maxL$ and $maxM$ respectively, $limL = 0.1092$ and $limM = 0.2127$. With such values, the deleted node X has the role of a *leader* if it belongs to the set $\{a, c, f, j\}$, the role of a *mediator* if it corresponds to e or f, and *other* if it belongs to the set $\{b, d, g, h, i, k, l\}$.

B. Proposed Approach

As mentioned in [3] and [4], 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(s) 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 will be sought.

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

1) Leader: If the deleted node L is a *leader*, i.e., it has a high degree centrality, then its links to other nodes are also deleted but the network should (if possible) exhibit at least one *leader*. That is why L is replaced with a new set of nodes L' which may contain one new *leader*, or nodes which separately are not *leaders*, but grouped together exhibit a leadership. Thereby, one finds L' and adds to L' enough links with some other nodes of the network to enforce the leadership of the elements in L'.

2) Mediator: If the deleted node M 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 but the past interactions between nodes or groups of nodes via the deleted node should in some way persist through a new selected mediator. That is why M is replaced by a new set of nodes M' which may contain one new *mediator*, or nodes which separately are not *mediators* but grouped together form a mediator. Thereby, one finds M' and adds to M' enough links with some other nodes of the network to establish the role of of mediation of the elements in M'.

When the deleted node is a *leader* or a *mediator* and the set of substitutes is not empty, each substitute is linked with some other nodes of the network according to one of three established link options as defined by Procedure *Link* (see explanation below). However, if the set of substitutes is empty, then some nodes are linked with some other ones according to one of three possible link options using Procedure *CreateLinks* (see below). Then, the new network structure and the new set (possibly empty) of *leaders* or *mediators* are returned.

3) Other: If the deleted node has no role, i.e., it is neither a leader nor a mediator, then its links are also deleted but no substitute is sought because the node is not enough important to identify a substitute and generate new links with already existing nodes.
Once the new network structure is established, if some nodes are completely isolated, i.e., they are not linked with any other node, they are then deleted.

C. 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 has a complexity[3] of $O(nm + n^2 \log n)$ and incorporates the two steps of our approach: the identification of the substitute(s) of the deleted node, and then the link management. Given a social network RS, a deleted node X, three parameters: α (for the substitutes), β (for the links) as well as lim (for the role), the main procedure returns both the new predicted network RS' (where X disappears) and the new leader(s) or mediator(s) of RS'. The parameter α represents the allowed relative deviation of the substitute indicator value from the indicator value of the deleted node. Using our illustrative example and $\alpha = 0.1$, the potential substitutes for the leader c are those having a degree centrality at least equal to $0.364 \times (1 - 0.1) = 0.327$ in the network deprived from the deleted node and its associated links. This is the case of node f (see Table III-a). The parameter β represents the ratio of a given measure (degree centrality, betweenness centrality or common neighbors) of a current node i that another node w needs to have in order to be linked to i.
The first executed instruction in the main procedure identifies the role of X (as defined in Subsection IV-A2) via the function *Role*(X, RS, lim) which returns either *Leader*, or *Mediator*, or *Other* (Line 1). When X is a *leader*, the indicator used is the degree centrality (C_D) (Lines 3-5). When X is a *mediator*, the appropriate indicator is the betweenness centrality (C_B) (Lines 6-8). For these two cases (Leader and Mediator), the procedure looks for a set of nodes NR to replace X via Procedure *Substitutes*$(X, RS, \alpha, Indic)$ (Line 9 and Algorithm 2). If NR contains elements, each substitute $nr \in NR$ is linked to some other nodes of the network via the function *Link*$(nr, RS, OpLink, \beta, X)$ (see Lines 10-13 and further explanations). Otherwise (i.e., no substitute is found), the procedure links nodes of the network via the function *CreateLinks*$(RS, OpCreateL, \beta, X)$ (Lines 14-16). Then, isolated nodes are deleted (Lines 18-20). Finally, the procedure computes the value of the two indicators for nodes of the predicted network RS' (Lines 21-22) and returns RS' as well as the sets L of leaders and M of mediators of that new network (Line 23).

Procedure *Substitutes*$(X, RS, \alpha, Indic)$ returns a set of nodes that replace the deleted node X according to the role of X. If $Indic = C_D$, *Substitutes* uses the degree centrality C_D to identify the *leaders*. If $Indic = C_B$, *Substitutes* uses the betweenness centrality C_B to determine the *mediators*.

Algorithm 2 describes the procedure *Substitutes*$(X, RS, \alpha, Indic)$. Given a social network RS, a deleted node X, a threshold α and an indicator $Indic$, the procedure returns a set of nodes as substitutes for X. For that purpose, the procedure computes the value of the indicator

[3]This complexity is explained by the dominant cost of the betweenness centrality calculation in a network having n nodes and m edges.

PARAMETER	MEANING
lim	Threshold for the role of a deleted node
$limL$	Possible value of lim in case of a leader
$limM$	Possible value of lim in case of a mediator
α	Relative deviation of the substitute indicator value from the indicator value of the deleted node
β	Ratio of a given measure (e.g., degree centrality) of a current node i that another node w needs to have in order to be linked to i
$OpLink$	Options $IMP, OLD, or NGB$ to link a substitute i to another node w
$OpCreateL$	Options $IMPC, OLD, or NGBC$ to link nodes when no substitute is found.

TABLE II
MEANING OF PARAMETERS.

Algorithm 1 PredictStruct(RS, X, α, β, $OpLink$, $OpCreateL$, lim)

Require:
$RS = \langle V, E\rangle$: initial network,
X: deleted node,
α, β, lim: thresholds,
$OpLink, OpCreateL$: parameters for possible options.
Ensure:
$RS' = \langle V', E'\rangle$: new network,
L, M: sets of nodes *Leader* and *Mediator* of RS'.

$r_X^{RS} \leftarrow$ *Role(X,RS,lim)*
if $r_X^{RS} \neq$ OTHER **then**
  if $r_X^{RS} =$ LEADER **then**
5: $Indic \leftarrow C_D$
  end if
  if $r_X^{RS} =$ MEDIATOR **then**
    $Indic \leftarrow C_B$
  end if
10: $NR \leftarrow$*Substitutes*$(X, RS, \alpha, Indic)$
  if $NR \neq \emptyset$ **then**
    for each $nr \in NR$ **do**
      $RS' \leftarrow$ *Link(nr, RS, OpLink, β, X)*
    end for
15: **else**
    $RS' \leftarrow$ *CreateLinks(RS, OpCreateL, β, X)*
  end if
end if
if $\exists y \in V' | \forall z \in V', z \neq y, \nexists e_{yz} \in E'$ **then**
20: $V' \leftarrow V' - \{y\}$ //deletion of isolated nodes
end if
$L \leftarrow \{Leaders\ of\ RS'\}$
$M \leftarrow \{Mediators\ of\ RS'\}$
return (RS', L, M)

$Indic$ of the nodes in the network $RS' = \langle V', E'\rangle$, i.e., without X and its associated links[4] (Lines 2-3) and stores the possible substitutes (in the set NR) having an indicator value in RS' close to (i.e., deviating by a relative proportion α from) the indicator value of X in RS (Lines 4-5). If NR contains more than one substitute, only one node z with the maximal indicator value (and possibly verifying some constraints) is returned as a substitute of X (Lines 8-10). If NR is empty (i.e., no substitute is found), then we compute a minimal group of nodes which has a sufficient indicator value to replace X via the function *Group*$(RS', \alpha, Indic)$ (see Lines 11-14). The set NR containing the substitutes of X is finally returned (Line 15).

[4]Since we try to predict the network structure after the deletion of one of its nodes, it seems realistic to focus on the impact of this deletion on the network deprived of the deleted node.

Algorithm 2 Substitutes(X, RS, α, $Indic$)

Require:
X: the deleted node,
$RS = \langle V, E\rangle$: initial network containing nodes and edges,
α: threshold,
$Indic$: degree or betweenness centrality indicator.
Ensure:
NR: set of potential substitutes for X

$RS' \leftarrow \langle V' = V - \{X\}, E' = E - \{e_{Xj}, \forall j \in V\}\rangle$
for each $w \in V'$ **do**
  Compute $ValIndic(w, RS', Indic)$
5: **if** $ValIndic(w, RS', Indic) \geq (1-\alpha) \times ValIndic(X, RS, Indic)$ **then**
    $NR \leftarrow NR \cup \{w\}$ //all possible substitutes
  end if
end for
if $|NR| > 1$ **then**
10: $NR \leftarrow SelectOne(\{z \in NR | ValIndic(z, RS', Indic)$ is maximal$\})$ //Select one node having the best $ValIndic$ value
end if
if $|NR| = 0$ **then**
  $NR \leftarrow$ *Group(RS',α,Indic)*
  //No substitute: set of nodes that form a group to replace X
15: **end if**
return NR

D. Additional Functions

Functions *Role*, *ValIndic*, *Group*, *Link* and *CreateLinks* can be briefly described as follows.

Role(i, R, lim) returns the role of a given node i within the network R by looking for the maximal values $maxL$ and $maxM$ among the network nodes and by comparing them to the threshold lim (as defined in Section IV-A2). In case i has two roles, it will be a *leader* if $\frac{C_D^R(i)}{maxL} \geq \frac{C_B^R(i)}{maxM}$. Otherwise, i will be a *mediator*.

ValIndic(i, R, Indic) returns the value of the indicator $Indic$ (either the degree centrality or the betweenness centrality) of the node i within the network R.

Group(R, α, $Indic$) returns the minimal subset G of V (where $R = \langle V, E\rangle$) such that the value of the indicator $Indic$ of G (as defined in Section IV-A1) has a relative deviation at most equal to α from the indicator value of the deleted node.

Link(i, R, OpLink, β, X) links the substitute node i in R 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 $C_D^{R'}(z)$ (resp. $C_B^{R'}(z)$) represents at least a proportion β of $C_D^{R'}(i)$ (resp. $C_B^{R'}(i)$) in R'. OLD links i to nodes which were associated with the deleted node X in the initial network R, in order to maintain the previous interactions. NGB links i to node z if the latter has a number of common neighbors with i greater than $\beta \times p$

	Degree centrality	Betweenness centrality
a	0.2	0.022
b	0.1	0
d	0.1	0
e	0.1	0
f	**0.4**	**0.333**
g	0.1	0
h	0.1	0
i	0.2	0.267
j	0.3	0.222
k	0.2	0
l	0.2	0

(a) After the deletion of c

	Degree centrality	Betweenness centrality
a	0.3	0.011
b	0.2	0
d	0.2	0
e	0.1	0
f	**0.7**	**0.811**
g	0.1	0
h	0.1	0
i	0.2	0.467
j	0.3	0.356
k	0.2	0
l	0.2	0

(b) After adding new links to f

TABLE III
INDICATOR VALUES FOR RS'.

where p is the maximum number of common neighbors that i shares with other nodes.

CreateLinks(R', $OpCreateL$, β, X) is used when no substitute is found and aims to create links between nodes which were connected to the deleted node X. The kind of the link depends on the selected option of $OpCreateL$ which can be $CLIQUE$, $IMPC$ or $NGBC$. The option $CLIQUE$ forms a clique with the identified nodes while the option $IMPC$ (resp. $NGBC$) has the same meaning as IMP (resp. NGB) associated with the function *Link*.

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 with 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 more realistic options for creating links between nodes are $NGBC$ and $IMPC$ and will more likely lead to a completely disconnected graph where every node will disappear.

As an illustration, let us consider the network RS given in Section II, Figure 1-a and Table I and let apply Procedure *PredictStruct*(RS, X, α, β, $OpLink$, $OpCreateL$, lim) with the following values: $X = c$, $\alpha = 0.1$, $\beta = 0.9$, $OpLink = OLD$, $OpCreateL = CLIQUE$, and $lim = 0.1$. The node c to delete has the role of a leader since $0.364 - C_D^{RS}(c) = 0$ which is less than 0.1. The variable $Indic$ will then take the value C_D (degree centrality). When Procedure *Substitutes*($c, RS, 0.1, C_D$) is called, node c and its associated links are first deleted from RS (see Figure 1-b). Then, node f is selected as the only possible substitute (see Table III-a) because its degree centrality in RS' is equal to 0.40 which is greater than $0.364 \times (1 - 0.1) = 0.327$. When Procedure *Link*($nr, RS, OpLink, \beta, X$) is called with the following values: $nr = f$, $OpLink = OLD$, $\beta = 0.9$ and $X = c$, we get the network RS' displayed in Figure 2-a where f is newly linked to the nodes that were attached to the deleted node c, i.e., nodes a, b and d. The *leader* of the new network is f while the *mediator* is also f (see Table III-b).

E. Improvement

1) Intuition: 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.

Consider the network RS given in Section II and Figure 1-a where we assume that three communities exist: $C_1 = \{a, b, c, d\}$, $C_2 = \{f, g, h\}$, and $C_3 = \{j, k, l\}$. The substitutes of c can be searched within the community of c, i.e., C_1. In that case, a possible substitute is a and a possible predicted network is the one of Figure 2-b. The substitutes of c can alternately be sought outside the community that contains c, like C_2 for example. In that case, a possible substitute is f and the predicted network is the one of Figure 2-a.

2) Communities: The detection of communities in a network is an important issue in social network analysis [11] 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 [12] which avoids having isolated nodes but has high computational requirements. Another commonly used method is based on the modularity maximization [13] 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 [14]. Recent studies focus on community analysis and their evolution. For instance, [15] takes into account the known communities at time t to determine the communities at time $t + 1$.

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 [16] which also allows the construction of a smaller comprehensible structure.

V. EXPERIMENTS

In this section, we empirically evaluate the potential of our approach for predicting the structure of the network when a node disappears. 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.

A. The datasets

We use two commonly known and large undirected and unweighted networks that we respectively call $COAUTHOR$ and $POWER\ GRID$[5]

The dataset $COAUTHOR$ is a co-authorship network of scientists working on network theory and experiment, as established by Newman [17]. It contains 1589 nodes and 2742 links. The dataset $POWER\ GRID$ is the largest downloadable network that 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 $COAUTHOR$.

B. Performance Analysis

The experimentations 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 each possible node role (*Leader, Mediator, Other*), for each possible option (*OLD, IMP, or NGB*) when substitutes exist ("WITH substitutes" on Figure 3) and for each possible option (*CLIQUE, IMPC, or NGBC*) when no substitute exists ("WITHOUT substitutes" on Figure 3). The overall mean time for the six options and the three roles is given through the "WITH and WITHOUT" chart[6] on Figure 3. Moreover, knowing that *Leader* and *Mediator* are similarly treated within the algorithm, performance results show that they perform similarly. Thus, no differentiation between the two 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.

Figure 3 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. In fact, when substitutes exist, the system processes only these nodes. However, when no substitute exists, the system handles all the nodes of the network. Note that the high execution times for the $POWER\ GRID$ dataset in Figure 3 are mainly due to the option $NGBC$ which computes the common neighbors of two nodes in a network. However, the merit of the option $NGBC$ is its ability to provide a good prediction based on the common neighbor measure as stated in [5].

C. Substitute Quality Analysis

To evaluate the quality of the prediction for substitutes, we use the classical recall and precision measures as well as the F-measure which are defined as follows:

$$Recall = \frac{|S_{rel} \cap S_{retri}|}{|S_{rel}|}\ ,\ Precision = \frac{|S_{rel} \cap S_{retri}|}{|S_{retri}|}$$
$$F\text{-}measure = \frac{2 \times (recall \times precision)}{(recall + precision)}$$

[5]Available at http://www-personal.umich.edu/ mejn/netdata/

[6]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 whose role is *Other*.

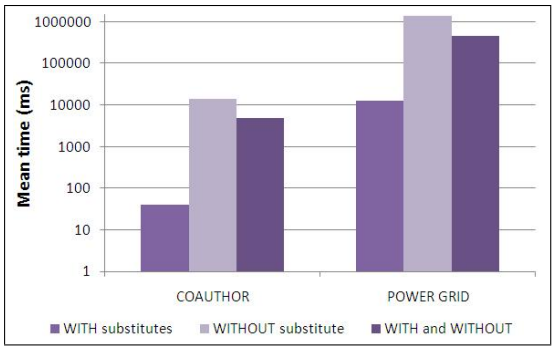


Fig. 3. Mean time (in milliseconds) when substitutes exist or not.

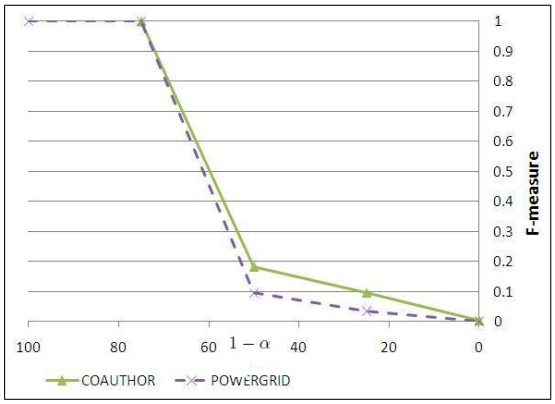


Fig. 4. F-measure of the substitutes for the two datasets according to α

where S_{rel} is the set of relevant substitutes and S_{retri} is the set of substitutes retrieved by our main procedure.

In other tests we have noticed that the parameter α, used to determine the substitutes, influences the number of (i) possible substitutes, (ii) added/deleted links and (iii) isolated nodes. As one can expect, substitute and link numbers increase when α increases (and conversely the smaller α is, the lower is the number of substitutes and links). The number of isolated nodes slightly decreases when α decreases. This can be explained by the fact that the number of substitutes decreases, which leads to a reduced number of isolated nodes.

Figure 4 displays the recorded F-measure for each dataset according to the value of α. It shows that the curves have a classical appearance and that smaller the substitute search interval is (i.e., the higher is $1-\alpha$), the better is the F-measure. For the two datasets, the F-measure is higher than 0.8 for $1 - \alpha \geq 70\%$, which is a good indication of the precision provided by our approach. The absence of values for $COAUTHOR$ between 100% and 80% is due to the absence of substitutes.

D. Prediction Quality Analysis

To evaluate the quality of our predicted network, we also use other features. Indeed, if we seek a predicted network close to the initial network, in terms of node role and network density, then we have to look for the mean gain (or loss) of density and indicator values (centrality degree, C_D and betweenness centrality, C_B) after restructuring the network. In fact, a mean gain (or loss) close to 0 is sought because we seek "a return to normal" following the deletion of one of the nodes in the

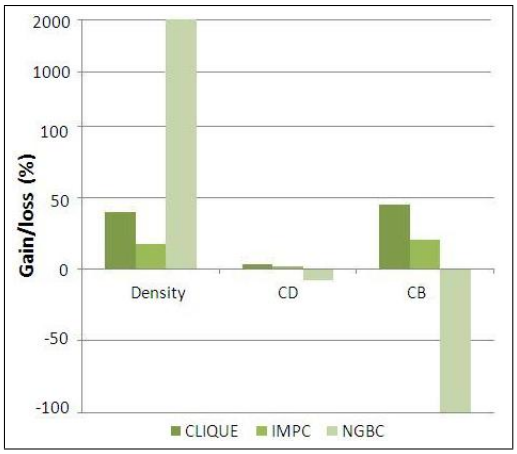


Fig. 5. Prediction quality: $COAUTHOR$, $\alpha = 0\%$

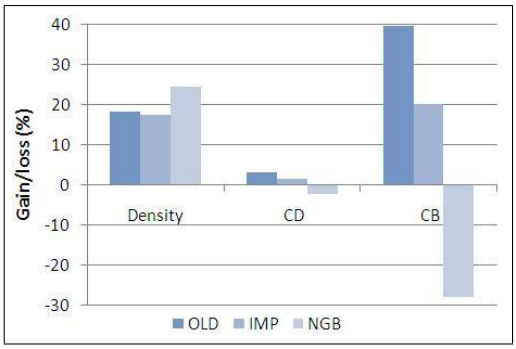


Fig. 6. Prediction quality: $COAUTHOR$, $\alpha = 20\%$

social network. Figures 5 and 6 display the tests conducted on $COAUTHOR$ dataset when substitutes exist ($\alpha = 20\%$) or not ($\alpha = 0\%$) and show that the best results are obtained with IMP and $IMPC$. Option $NGBC$ does not seem effective because the gain of density is higher than 2000% and the mean loss of betweenness centrality reaches 100%.

To the best of our knowledge, there is no study that handles the same issue as ours, except the work in [7] 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.

VI. 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* or *mediator*) played by entities in terms of their interactions within 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 IMP and $IMPC$ offer relatively good results in terms of execution time and precision (quality of prediction). However, option $NGBC$ is the worst both in terms of execution time and precision.

Our future work includes a linkage between our approach and the role transfer problem as analyzed in [18]. Moreover, we plan to propose a substitution method that could use other measures (or scores) instead of centrality or betweenness degree to further use a probabilistic approach such as Dirichlet mixtures or Bayesian networks. The work can also be extended to handle the deletion of more than one node at a time.

Acknowledgment

We acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and would like to warmly thank referees for their comments and suggestions that helped improve the quality of this paper.

References

[1] P. J. Carrington, *Models and methods in social network analysis*, ser. Structural analysis in the social sciences. Cambridge Univ. Press, 2005.
[2] M. Jamali and H. Abolhassani, "Different aspects of social network analysis," in *IEEE/WIC/ACM International Conference on Web Intelligence*, 2006, pp. 66–72.
[3] E. Ben-Naim and P. L. Krapivsky, "Addition - deletion networks," *Journal of Physics A: Mathematical and Theoretical*, vol. 40, no. 30, p. 8607, 2007.
[4] R. Toivonen, L. Kovanen, M. Kivel, J.-P. Onnela, J. Saramki, and K. Kaski, "A comparative study of social network models: Network evolution models and nodal attribute models," *Social Networks*, vol. 31, no. 4, pp. 240–254, October 2009.
[5] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*. New York, NY, USA: ACM, 2003, pp. 556–559.
[6] T. Tylenda, R. Angelova, and S. Bedathur, "Towards time-aware link prediction in evolving social networks," in *The 3rd SNA-KDD Workshop '09 (SNA-KDD'09)*. SIGKDD, June 2009.
[7] D. M. A. Hussain and Z. Ahmed, "Dynamical adaptation in terrorist cells/networks," in *SCSS (2)*, 2008, pp. 557–562.
[8] S. Wasserman and K. Faust, *Social network analysis : methods and applications*, 1st ed., ser. Structural analysis in the social sciences, 8. Cambridge University Press, November 1994.
[9] M. G. Everett and S. P. Borgatti, *Models and methods in social network analysis*. Cambridge University Press, 2005, pp. 57,76.
[10] U. Brandes, "A faster algorithm for betweenness centrality," *Journal of Mathematical Sociology*, vol. 25, pp. 163–177, 2001.
[11] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, February 2010.
[12] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.
[13] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Phys. Rev. E*, vol. 69, no. 6, p. 066133, Jun 2004.
[14] J. Chen, O. R. Zaïane, and R. Goebel, "Detecting communities in social networks using max-min modularity," in *SDM*, 2009, pp. 978–989.
[15] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng, "Analyzing communities and their evolutions in dynamic social networks," *ACM Trans. Knowl. Discov. Data*, vol. 3, pp. 8:1–8:31, April 2009.
[16] H. C. White, S. A. Boorman, and R. L. Breiger, "Social structure from multiple networks: I. blockmodels of roles and positions," *American Journal of Sociology*, vol. 81, pp. 730–780, 1976.
[17] M. E. J. Newman, "Coauthorship networks and patterns of scientific collaboration," *Physical Review E*, vol. 74, p. 036104, 2006.
[18] H. Zhu and M. Zhou, "M-m role-transfer problems and their solutions," *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, vol. 39, no. 2, pp. 448–459, 2009.