PeerJ Preprints NOT PEER-REVIEWED

A Mixed Graph Framework to evaluate the complementarity of communication Tools

Leonardo Carvalho [1] , **Eduardo Bezerra** [1] , **Gustavo Guedes** [1] , **Laura Assis** [1] , **Leonardo Lima** [1] , **Rafael Barbastefano** [1] , **Artur Ziviani** [2] , **Fabio Porto** [2] , **Eduardo Ogasawara** [Corresp. 1]

[1] CEFET/RJ, Brazil
[2] LNCC, Brazil

Corresponding Author: Eduardo Ogasawara
Email address: eogasawara@ieee.org

Due to the constant innovations in communications tools, several organizations are constantly evaluating the adoption of new communication tools (NCT) with respect to current ones. Especially, many organizations are interested in checking if NCT is really bringing benefits in their production process. We can state an important problem that tackles this interest as for how to identify when NCT is providing a significantly different complementary communication flow with respect to the current communication tools (CCT). This paper presents the Mixed Graph Framework (MGF) to address the problem of measuring the complementarity of a NCT in the scenario where some CCT is already established. We evaluated MGF using synthetic data that represents an enterprise social network (ESN) in the context of well-established e-mail communication tool. Our experiments observed that the MGF was able to identify whether a NCT produces significant changes in the overall communications according to some centrality measures.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017

PeerJ Preprints NOT PEER-REVIEWED

A Mixed Graph Framework to Evaluate the Complementarity of Communication Tools

Leonardo Carvalho[1], Laura Assis[2], Leonardo Lima[3], Eduardo Bezerra[4], Gustavo Guedes[5], Artur Ziviani[6], Fabio Porto[7], Rafael Barbastefano[8], and Eduardo Ogasawara[9]

[1]**CEFET/RJ**
[2]**CEFET/RJ**
[3]**CEFET/RJ**
[4]**CEFET/RJ**
[5]**CEFET/RJ**
[6]**LNCC**
[7]**LNCC**
[8]**CEFET/RJ**
[9]**CEFET/RJ**

Corresponding author:
Eduardo Ogasawara[1]

Email address: eogasawara@ieee.org

ABSTRACT

Due to the constant innovations in communications tools, several organizations are constantly evaluating the adoption of new communication tools (NCT) with respect to current ones. Especially, many organizations are interested in checking if NCT is really bringing benefits in their production process. We can state an important problem that tackles this interest as for how to identify when NCT is providing a significantly different complementary communication flow with respect to the current communication tools (CCT). This paper presents the Mixed Graph Framework (MGF) to address the problem of measuring the complementarity of a NCT in the scenario where some CCT is already established. We evaluated MGF using synthetic data that represents an enterprise social network (ESN) in the context of well-established e-mail communication tool. Our experiments observed that the MGF was able to identify whether a NCT produces significant changes in the overall communications according to some centrality measures.

1 INTRODUCTION

Communication tools are in constant evolution. They usually change the way people collaborate with each other. Not long ago, letters, telegrams, and other written communications on paper were the mainstream. However, since the beginning of the Internet, communication tools were extended through e-mail. The use of e-mail is widespread and almost ubiquitous in enterprises, being responsible for the majority of the communication flow inside them (Bennett, 2012).

Innovations in communication tools continue to occur and several new tools, such as instant messaging, blogs, and content management have been developed (Hansen et al., 2010). All these tools, when applied in the enterprise scene, target the increasing of productivity and collaboration among employees. Recently, new opportunities to empower communication among employees have arisen with the advent of online social networks (OSNs) (Raghavan, 2002). Although e-mail is adequate for certain types of communication, instant messaging (IM), wikis, and other social applications can be better options for collaborative work (Friedman et al., 2014) and are, thus, gaining momentum in enterprise communications.

As an emerging communication technology, OSNs provide a variety of communication services such as profiles, comments, private messaging, blogging, media file sharing, and instant messaging. Some of these communication tools provide their services through a mobile network (Chai and Kim, 2012). These features are important as they help breaking existing barriers to communication among collaborators,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017

regardless of their position in the organization chart. They can stimulate interactions involving employees that are far apart in an enterprise hierarchy (Friedman et al., 2014).

Due to demands of privacy and other strategic decisions, enterprises also may choose to establish private social networks that are restricted to employees and collaborators of their main business (Ning et al., 2012). These networks, known as enterprise social networks (ESNs), are commonly protected by firewalls and restricted to employees (Leftheriotis and Giannakos, 2014). These tools are commonly inspired on public social networks, such as Facebook, LinkedIn, including Web 2.0 collaboration tools (Turban et al., 2011).

There are some specialized enterprise social network, such as Connections (Zaffar and Ghazawneh, 2012) and Microsoft Sharepoint (Rooksby and Sommerville, 2011). Their usage, however, differs a little from traditional public social networks (Ning et al., 2012). In particular, they have to focus on enterprise issues, such as improvement of intra-enterprise communication and better integration with other enterprise tools.

Under this perspective and due to investments, enterprises are concerned to measure the effective adoption of a new communication tool (NCT). Particularly, these issues are relevant for Small Medium Enterprises (SMEs). They are thus searching for an effective way to assess if a NCT is really bringing benefits for their productive process. We can state the problem as how to identify when a NCT is providing a complementary communication flow with respect to the current communication tools (CCTs) that are being used.

In this paper, we address the problem of measuring the complementarity of a NCT in the scenario where some CCT is already an established tool. In order to do that, we present the Mixed Graph Framework (MGF), which is designed to evaluate how complementarity the involved communication tools are by using a mixed graph modeling. The proposed MGF is based on the premise that the CCT can be considered as a baseline for evaluating any other tool to improve communication in enterprises. It is important to use a common representation of the communication flow to enable comparison between them. In this work, the communication networks from the CCT and the NCT are modeled as graphs, named G_c and G_n, respectively. From these graphs, the MGF produces a mixed graph G_m to measure if a NCT is acting as a complementary tool among employees as compared with the CCT.

We have evaluated MGF using synthetic data that represents SME communication flows. In our experiment, we assume that e-mail is the CCT and an Enterprise Social Network (ESN) is the NCT. Based on the shared messages in both tools, we compute several metrics and conduct a statistical analysis on them to evaluate the complementarity of the NCT. Our experiments observed that the MGF was able to identify whether an NCT produces significant changes in the overall communication.

The remainder of the paper is organized as follows. Sections 2 and 3 present related work and the general background, respectively. The proposed MGF is described in Section 4. Section 5 presents our experimental evaluation. Finally, Section 6 concludes the paper.

2 RELATED WORK

The analysis of social networks is widely explored and it has been studied for several years (Ngai et al., 2015). Many of these studies focused on the information that can be extracted from these networks analyzing their dynamics and structure. When it comes to the impacts of communication tools adoption in an enterprise environment, the need of study in this area expanded in recent years (Friedman et al., 2014). These studies focus on the impacts of the usage of an ESN, and served as a basis for administrators to preview what challenges relating to this new trend can cause in the near future of the enterprise and how to use them in favor of business objectives.

One of the main concerns about the adoption of communication tools in an enterprise environment is related to the notion of being social. A common question is if social in this context means something connected only to interpersonal relationship. Wasko et al. (Wasko et al., 2009) showed that enterprise communication tools were used not only for the maintenance of an interpersonal relationship, but also to discuss about the core business.

A case study of adoption and implementation of ESN can be found in Cross et al. (Cross et al., 2001). The study undertook the mapping of information flow from executives in the exploration and production division of the British Petroleum (BP) Company. The work examined the adoption of social networking tools as a way to transfer and disseminate knowledge. The analysis of the communication flow between twenty managers of exploration and production area revealed a striking contrast between the structures of

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017

formal and informal groups. Although BP has a strong hierarchical and functional structure, the study showed the great importance middle managers have in general communication. The study also showed that these middle managers were critical to maintain the information flow between areas.

We can find several publications on the use of open social networks and e-mail, most of them related to representation of information flow as a graph. There are several works that apply graph theory to social network analysis. Johnson et al. (Johnson et al., 2012) studied the communication and friendship relationships extracted from e-mail data. The study analyzed aspects related to the distribution of groups and centrality, the authors investigated the growth of the corresponding e-mail network.

Once a social network is represented as a graph, it is possible to extract metrics that enables data mining (Nettleton, 2013), searching for domain experts in an ESN (Chen et al., 2006), and identifying groups (clusters or cliques) that are related to concentrations of communication flows inside the graph (Prado and Baranauskas, 2013). Many of these metrics, such as cohesion and average distance, are useful in network analysis, as they enable insights about how communication flows in a network and the proposal of improvements (Newman, 2003).

In Hamulic et al. (Hamulic and Bijedic, 2009), the authors structure and compare social networks by analyzing the communication flow among students of courses available in a Distance Learning scenario. Their study showed that these social networks could be used to analyze the communication flow and draw conclusions to improve the available e-learning courses.

In addition to classic social network analysis methods and its variations, Stewart et al. (Stewart and Abidi, 2012) show how data visualization and statistical analyses provide a broad view of the communication patterns within the discussion forums. They show how such analyses relate the general behavior of the social network, isolating potential core group members of the social network and exploring existing intergroup relations between institutions and professions.

With the growing popularity of online social networks, it is appealing to develop social network frameworks (SNF) for a variety of environments to improve communication. In fact, some studies using a framework for analysing social networks can be found in the literature. Turban et al. (Turban et al., 2011) adopt the fit-viability framework to deal with the adoption of social networks for specific tasks or projects. Lynn et al. (Lynn et al., 2015) proposed a general framework for researchers to understand and analyse social media using big data. The big data in this case arises from the relationships between entities within a social network sites. The proposed framework accommodates different data types and methods. Kim et al. (Kim et al., 2013) suggest an evolutionary framework for analyzing the intergenerational transition of Online Social Networks.

Some researchers have proposed frameworks for understanding social media and guiding research such as Chai et al. (Chai and Kim, 2012) that suggest a theoretical framework to understand social networking site users' knowledge contribution behavior. A causal-chain framework was developed by Ngai et al. (Ngai et al., 2015) in order to understand the inter-relationships among different research constructs adopted.

There are many works that analyse social networks, and study its behavior. Also, there are some papers that propose frameworks for these purposes. Nonetheless, as far as we know there is no other work that proposes and implements a Mixed Graph framework to measure if a NCT is being complementary to a CCT already in use.

3 BACKGROUND

This section presents the fundamental concepts used in our framework, and is organized in three main subsections. Section 3.1 presents general graph concepts. Section 3.2 describes the major centrality-based measures that are used as input for the performed statistical analysis. Section 3.3 presents the general statistical tests for non-parametric data sets.

3.1 Graph Representation

Using graph theory terminology (see Ahuja et al. (Ahuja et al., 1993) for details), communication networks (such as e-mail and ESN) can be modelled as a weighted directed graph $G(V,E)$, where V is the set of $|V|$ nodes and E is the set of $|E|$ edges. A node $i \in V$ represents a collaborator with a connection point. The arcs $(i, j) \in E$, $i \in V$ and $j \in V$ represent a communication link between two collaborators (Figure 1).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017

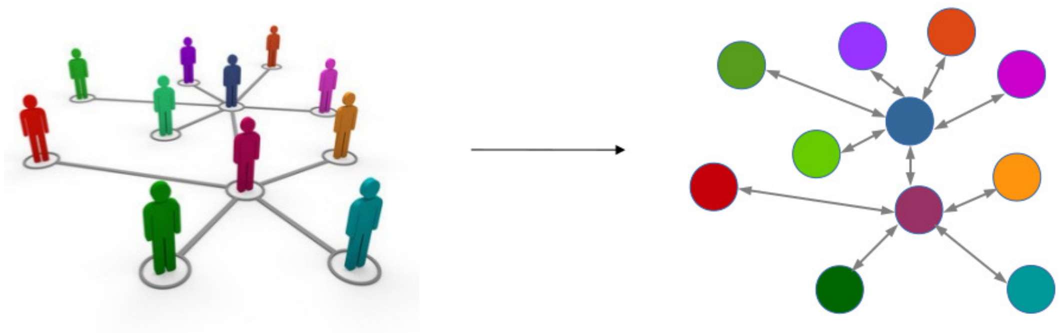


Figure 1. Graph representation of a NCT or a CCT network

A weight $w_{ij} > 0$ is assigned to each edge with ending nodes i and j and represents the amount of communication flow between these two nodes. Since $G(V,E)$ is directed, it may be that $w_{ij} \neq w_{ji}$. The adjacency matrix $A_{i,j} = a_{i,j}$ of the weighted graph G can be defined as:

$$a_{ij} = \begin{cases} w_{ij}, & \text{if there is an edge connecting the node } i \text{ to } j, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

3.2 Graph Centrality Measures

When some problem is modelled by a graph, many properties are associated with each node, such as distance and centrality. These properties provide a summary of the graph.

A widespread centrality measure is the weighted closeness of a node v (Opsahl et al., 2010). If a node v represents a collaborator in an enterprise network, the closeness of v measures how close a collaborator is to others. Collaborators that occupy central positions with respect to closeness are important in communication (Wasserman and Faust, 1994). The weighted closeness of a node v is computed by

$$C_c(v) = \frac{1}{\sum_{x \in V \setminus v} d(v,x)}, \tag{2}$$

where $d(v,x)$ is the weighted geodesic distance between the nodes v and x.

Another well-known measure is the weighted betweenness centrality of a node v (Kolaczyk, 2009). It is a measure aimed at summarizing the extent to which a vertex is located 'between' other pairs of vertices. Let us introduce some notation before formally define the betweenness centrality. Consider arbitrary nodes $u, v \in V$. A path $P(u,v)$ which starts at u and finishes at v is an ordered sequence of nodes, $P(u,v) \ = < u \ = \ v_1, v_2, \ldots, v_k \ = \ v >$, such that $e_i = (v_i, v_{i+1}) \in E$ for $i = 1, \ldots, k-1$. The length of the path $P(u,v)$ is given as the sum of the edge weights of the path and the shortest path function $s_G(u,v)$ between nodes $u, v \in V$ is given by

$$s_G(u,v) = \min_{P(u,v)} \sum_{i=1}^{k-1} w_{i,i+1}.$$

The betweenness centrality for any given node $v \in V$ is then given by

$$C_b(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}, \tag{3}$$

where $\sigma(s,t)$ is the number of paths $P(s,t)$ of size $s_G(s,t)$ connecting s and t and $\sigma(s,t|v)$ is the number of shortest paths passing through vertex v.

A third class of centrality measure is the *kleinberg centrality* introduced by (Kleinberg, 1998). The main idea is to identify good relative *hubs* and *authorities*' nodes. A hub is a node that points to many important nodes, and an authority node is the one that is pointed by many important nodes. Both are based on the eigenvectors related to the largest eigenvalues of the matrices AA^T and A^TA. The hub centrality of

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017

the node v_i, denoted here by $C_h(v_i)$, is the $i-$th entry of the vector x satisfying Equation (4), where $\lambda \in \Re$ is the largest eigenvalue of AA^T.

$$AA^T x = \lambda x. \tag{4}$$

Similarly, the authority of a node v_i, denoted here by $C_a(v_i)$, is the $i-$th entry of the vector y satisfying Equation (5), where $\beta \in \Re$ is the largest eigenvalue of $A^T A$.

$$A^T Ay = \beta y. \tag{5}$$

3.3 Statistical Analysis

The need to compare two different data sets is very common. Such comparison may vary according to the objectives of study. We can summarize two different statistical tests that are relevant to compare two data sets: (i) distribution; and (ii) correlation (Larsen and Marx, 2005). For each one of these scenarios, there is a set of statistical tests that can be used. They vary according to the distribution of the data sets. Commonly, social medias are scale-free networks and follow a power-law distribution. In this case, non-parametric tests are more adequate. For the sake of simplicity, we are going to present one statistical test for comparing two data sets.

Mann-Whitney U test, also known as Wilcoxon rank sum test from the difference in medians, is a distribution analysis test. The goal of this test is to measure the extent to which the medians of two independent data sets are different from each other, *i.e.*, to check if the difference between the median of these two data sets is significantly different from zero.

Spearman rank correlation test is a correlation analysis test, whose goal is to test if the rank correlation coefficient between two variables is significantly different from zero. The null hypothesis establishes zero correlation between two variables.

4 THE MIXED GRAPH FRAMEWORK (MGF)

In this section, we present a framework to evaluate the complementarity of communication tools using a mixed graph modeling, called here Mixed Graph Framework (MGF). Algorithm 1 summarizes how the MGF works. The first two lines (2-3) are related to modeling graphs for communication tools and are described in further detail in Section 4.1. Line (4) is described in Section 4.2 and produces the mixed graph. Line (5) is described in Section 4.3 and computes centrality measures to evaluate the complementarity of the NCT with respect to the CCT.

Algorithm 1 Main *MGF* Algorithm

1: **function** $MGF(D\ d_c, D\ d_n, ef_c, ef_n)$
2: $\quad G_c \leftarrow fExtract_c(d_c)$
3: $\quad G_n \leftarrow fExtract_n(d_n)$
4: $\quad G_m \leftarrow fMix(G_c,\ G_n)$
5: $\quad$ return $fAnalyze(G_c,\ G_m)$
6: **end function**

1: **function** $fAnalyze(G_c,\ G_m)$
2: $\quad r_1 \leftarrow analyzeClosenessDistribution(G_c,\ G_m)$
3: $\quad r_2 \leftarrow analyzeClosenessCorrelation(G_c,\ G_m)$
4: $\quad r_3 \leftarrow analyzeBetweennessCorrelation(G_c,\ G_m)$
5: $\quad r_4 \leftarrow analyzeEigenTopK(G_c,\ G_m)$
6: $\quad$ return $\{r_1, r_2, r_3, r_4\}$
7: **end function**

4.1 Extract Functions

The first two activities of Algorithm 1 encompass modeling graphs from the communication tools. Graphs $G_c = (V_c, E_c)$ and $G_n = (V_n, E_n)$ are, respectively, generated through the extraction Functions $fExtract_c$ and $fExtract_n$ that are applied over the CCT and NCT datasets.

A node $i \in V_c$ and $p \in V_n$ corresponds to collaborators of their respective graphs G_c and G_n. An edge $e_{i,j} \in E_c$ represents a communication in CCT from collaborator $i \in V_c$ to collaborator $j \in V_c$ and the edge

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017

weight $w_c(i,j)$ corresponds to the amount of messages exchanged from i to j. Similarly, an edge $e_{i,j} \in E_n$ represents a communication in the NCT from collaborator i to collaborator $j \in V_n$ and the edge weight $w_n(i,j)$ corresponds to the amount of messages exchanged from i to j.

Both $fExtract_c$ and $fExtract_n$ are User Defined Functions (UDFs) that vary according to the adopted communication tools. For example, if CCT is an e-mail tool, the communication flow in the graph G_c between two collaborators i and $j \in V$ is measured by the amount of email messages exchanged by them, as described in Equation (6). On the other hand, if the NCT is an ESN tool, the communication flow is measured by the weighted average of posts, comments, and likes someone is interested to extract from the ESN, as described in Equation (7).

$$w_c(i,j) = |mails(i,j)| \tag{6}$$

$$w_n(i,j) = \frac{\alpha|posts(i,j)| \ + \beta|comments(i,j)| \ + \gamma|likes(i,j)|}{\alpha+\beta+\gamma} \tag{7}$$

Figures 2(a) and 2(b) display illustrative examples of G_c and G_n, respectively. The graph G_c is obtained by applying $fExtract_c$ over the D_c dataset and the graph G_n is obtained by applying $fExtract_n$ over D_n dataset.

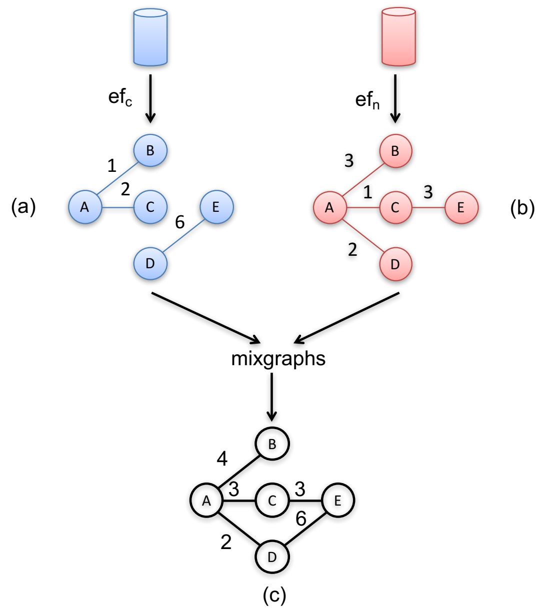


Figure 2. Communication flow: (a) G_c extracted from the CCT dataset; (b) G_n extracted from NCT dataset; (c) G_m produced by mixing G_c with G_n

4.2 Mixed Graphs

Let $G_m = (V_m, E_m)$ be the mixed graph with node set $V_m = V_c = V_n$ of order $|V_m|$ and edge set $E_m = E_c \cup E_n$. To each edge $e_{i,j} \in E_m$ a weight $w_m(i,j)$ is assigned as given by Equation (8). The mixed graph activity is described in Algorithm 2. It receives both G_c and G_n as an input and builds the mixed graph G_m of order $|V_m|$ with its edges weights given by the vector w_m.

$$w_m(i,j) = w_c(i,j) + w_n(i,j) \tag{8}$$

Note that the graph G_m represents the total flow of communication provided by the two communication tools and can be used to identify whether the NCT is actually changing the communication flow or just

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017

mirroring existing communication flows between collaborators in the CCT. An example of G_m can be observed in Figure 2(c) obtained from G_n and G_c.

Algorithm 2 Mixed Graphs

```
function fMix(G_c, nG)
    V_m ← V_c ∪ V_n
    G_m ← EmptyGraph(|V_m|)
    for i ← 1 to |V_m| do
        for j ← 1 to |V_m| do
            if i <> j then
                w_m(i, j) ← w_c(i, j) + w_n(i, j)
            end if
        end for
    end for
    return (mG, w_m)
end function
```

4.3 Complementarity Analysis

The complementarity analysis computes centrality measures of each vertex extracted from G_c and G_m. These values are used to compute if such metrics from G_c are statistically significant different from G_m. In this case, it indicates that G_n is not simply an overlap of G_c, *i.e.*, actually bringing complementarity in the overall communication. Such an activity is described in Algorithm 3.

It is worth mentioning that all centrality-based measures expect a weighted adjacency matrix as an input. However, in all built graphs (G_c, G_n, and G_m), the weight of the edges corresponds to the communication flow over a period. In this way, prior to any centrality computation, it is important to convert flows to distances since more messages, e-mails, and post exchanges imply less distance between two collaborators. Such a transformation is described by Function $convertDist(w)$ that applies Equation (9) for all edges in Algorithm 3.

$$\overline{w}(i, j) = \frac{1}{w(i, j)} \tag{9}$$

Functions closeness, betweenness, and Eigen, respectively compute the weighted closeness, weighted betweenness, and weighted Eigen vectors measures (Opsahl et al., 2010) of G_c and G_m. The first line in all functions described in Algorithm 3 is to convert the communication-based graph into a distance-based graph according to Equation (9).

Function $analyzeClosenessDistribution$ analyzes the closeness centrality distribution. The goals is to compute if the difference in the median of the closeness of each graph is significantly different from zero. For that, the nonparametric *Wilcoxon rank sum* is used (Devore and Berk, 2011). The intuition of this function is to compute if the introduction of NCT changes the amount of communication flow significantly with respect to the CCT.

Functions $analyzeBetweennessCorrelation$ and $analyzeClosenessCorrelation$ correlate the betweenness and the closeness centralities between G_c and G_m, respectively. For that, the nonparametric *Spearman correlation test* is used (Devore and Berk, 2011). The intuition of these functions is to compute if the introduction of the NCT changes significantly the way people interact with respect to the CCT by analyzing the established communication flows. This indicates if the NCT is not simply increasing the scale of messages among persons, but if it is changing the communication flow structure. Such a test is complementary to $analyzeClosenessDistribution$. We can have situations where $analyzeClosenessDistribution$ may not differ but either $analyzeClosenessCorrelation$ or $analyzeBetweennessCorrelation$ may present significant changes and vice-versa.

Function $analyzeHub$ is also a complementary analysis. It analyzes the influence of introducing new edges in the communication flow. It starts by multiplying the adjacency matrix with its transpose targeting the main hubs in the communication flow. This is done in both graphs (G_c and G_m). Inside the function, we compute the top-k more central vertices in both graphs and the overlap between them (same central

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017

Algorithm 3 Analysis of Centrality (Closeness, betweenness, Eigen)

function $analyzeClosenessDistribution(G_c,\ G_m)$
  $vc_c \leftarrow closeness(convertDist(G_c))$
  $vc_m \leftarrow closeness(convertDist(G_m))$
  return $wilcox.test(vc_m, vc_m, conf.level = 0.95)$
end function

function $analyzeClosenessCorrelation(G_c,\ G_m)$
  $vc_c \leftarrow closeness(convertDist(G_c))$
  $vc_m \leftarrow closeness(convertDist(G_m))$
  return $spearman.cor.test(vc_m, vc_m, conf.level = 0.95)$
end function

function $analyzeBetweennessCorrelation(G_c,\ G_m)$
  $vb_c \leftarrow betweenness(convertDist(G_c))$
  $vb_m \leftarrow betweenness(convertDist(G_m))$
  return $spearman.cor.test(vb_c, vb_m, conf.level = 0.95)$
end function

function $analyzeHub(G_c,\ G_m, k)$
  $ve_c \leftarrow eigen(asHub(convertDist(G_c)))$
  $ve_m \leftarrow eigen(asHub(convertDist(G_m)))$
  $ratio \leftarrow overlap(top_k(ve_c), top_k(ve_m))$
  $sig \leftarrow hypergeo(ratio, m = k \cdot |ve_c|, n = (1-k) \cdot |ve_c|)$
  return $\{ratio, sig\}$
end function

vertices in both graphs). We also compute the probability using hypergeometric distribution of such an occurrence.

The MGF is implemented in R and is made publicly available at sourceforge.[1] Statistical tests *Wilcoxon rank sum* and *Spearman correlation* test are available in many statistical packages, such as R (Dalgaard, 2008) and were included in *MGF*.

5 EXPERIMENTAL EVALUATION

This section presents the evaluation of the proposed MGF in measuring if the NCT brings complementarity to the CCT inside a Small Medium Enterprises (SME) (Hoffmann and Schlosser, 2001). We used synthetic data to simulate both CCT and NCT usage to explore the MGF under different group configurations and enterprise scales. Both MGF and experimental evaluation is made available at *https://github.com/eogasawara/mgf*.

We have organized this section in three parts, as follows. Section 5.1 discusses synthetic data preparation that models SME (Newman et al., 2002). In Section 5.2, we describe the general procedure of growth network used in experimental evaluation. In Section 5.3, we present a toy sample analysis to illustrate the benefits of MGF. In Section 5.4, we conduct a sensitive analysis of MGF under different SME scenarios.

5.1 Synthetic data generation

Many networks can be framed in the definition of scale-free networks (Barabási and Albert, 1999). A network is classified as scale-free if the degree distribution of its nodes follows the power law model (Newman et al., 2002). The network formed by the flow of messages sent within the CCT can also be classified as scale-free (Ebel et al., 2002). Scale-free networks have two general concepts: growth and preferential attachment. The concept of growth points out to the constant growth of the number of nodes in the network. The preferential attachment means that the more connected is a node, the more likely is that it gets new links. The basic understanding for this second concept is that a new member on the network has a higher probability to interact with a person who interacts with many people than with someone who is not so active in the network.

[1]https://sourceforge.net/p/gpca/wiki/MGF

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017

The most notable feature of a scale-free network is the existence of nodes with degree much higher than the average degree in the network. The highest degree nodes are often called hubs and have specific meanings in each network. The presence of hubs is directly related to the robustness of the network. Most of the nodes are not hubs, and the probability of a significant impact on total flow with the departure of one of these low degree nodes is very low. On the other hand, the removal of a hub can cause a large impact in the communication flow or even a network partition.

In the experiments presented in our work, we generated G_c (simulating e-mail communication) and G_n (simulating ESN communication) as scale-free networks. However, G_c follows the organizational structure (hierarchy of the network) formed by the e-mail information (Johnson et al., 2012), whereas the ESN does not impose such a constraint. This assumption is reasonable since most companies and institutions are organized hierarchically, with people in a group, reporting to a single manager who in turn reports to another manager (Hansen et al., 2010). In each group, people often connect directly to others without passing messages up and down the chain of command, but inter-group typically follows hierarchical structure evidenced in some e-mail communication studies (Wang et al., 2011).

Algorithm 4 generates synthetic instances of CCT and NCT; and was implemented using *poweRlaw*, an R package to create scale-free graphs. Initially, the first three parameters k, v, e are related to generation of the subgraphs that will form CCT graph (i.e., G_c). It starts by creating k subgraphs in G_c. Each subgraph has v nodes with e edges. After that, the most central nodes in each subgraph, according to its closeness centrality, are connected to each other to establish a hierarchical communication in G_c. In the end of G_c build phase, this graph has $|V_c| = v_c = v \cdot k$ nodes and $|E_c| = e_c = (e \cdot k) + k$ edges. Then, the NCT graph G_n is generated with $v_n = |V_n|$ nodes and $e_n = |E_n|$ edges, such that $v_n = v_c$. By construction, G_n is strictly scale-free.

Table 1. Parameters used in the experimental evaluation

Parameter	Description
$v_c = v_n$	Number of nodes in both graphs, G_c and G_n
k	Number of groups in G_c
e_c	Number of edges (communication flows) in G_c
e_n	Number of edges (communication flows) in G_n

Algorithm 4 Synthetic dataset production

```
1:  function SyntheticDatasets(k, v, e, e_n)
2:      for all i ← 1 to k do
3:          G_c^i ← new ScaleFreeGraph(v, e)
4:          G_c ← G_c ∪ G_c^i
5:      end for
6:      for i ← 1 to k_E − 1 do
7:          for j ← i + 1 to k_E do
8:              e_l ← connect(G_c^i, G_c^j)
9:              E_c ← E_c ∪ e
10:         end for
11:     end for
12:     v_c ← v · k
13:     v_n ← v_c
14:     G_n ← new ScaleFreeGraph(v_n, e_n)
15:     return ({G_c, G_n})
16: end function
```

In Section 5.4, we explore three scenarios produced during synthetic data generation that correspond to representative contexts for SME (Hoffmann and Schlosser, 2001; Eurostat, 2016), such as the number of vertices. In Europe (Eurostat, 2016), a small enterprise has a number of collaborators greater than 10 and lower than 50, whereas in medium enterprises the number of collaborators is greater than or equal to 50 and lower than 250. Additionally, the number of messages and edges explored in our study are

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017

in agreement with communications using both e-mail (Layman et al., 2006) and online social networks (Benevenuto et al., 2009). The scenarios adopted for SMEs are presented in Table 2.

Table 2. SME Scenarios

Scenario	Description
SE (G_n scale)	$v_c = 30,\ k_c = 3,\ e_c = 60$ $small:\ e_n = 25$ $medium:\ e_n = 45$ $large:\ e_n = 55$
SE (G_c groups)	$v_c = 30,\ e_c = 60,\ e_n = 45$ $low:\ k_c = 2$ $moderated:\ k_c = 3$ $high:\ k_c = 5$
ME (G_c groups)	$v_c = 150,\ e_c = 60$ $low:\ k_c = 10,\ e_n = 120$ $moderated:\ k_c = 15,\ e_n = 180$ $high:\ k_c = 25,\ e_n = 300$

5.2 Network Growth

Consider both G_c and G_n produced during the synthetic dataset production. We can apply the MGF to compute metrics and check if G_n is being complementary to G_c. However, to better explore MGF, in all experimental evaluation we conducted the analysis with G_n using a network growth described in Algorithm 5. The goal is to allow for the comprehension of the MGF behavior as we increase G_n from an empty graph until reaching the entire G_n structure. According to Algorithm 5, the growth ratio δ filter both edge weights and the number of edges in its entire structure according to its weight distribution. The edge weights for w_n are all multiplied by $\frac{\delta}{100}$, in order to set to relative strength of usage in both networks. The lesser the value of δ, the lesser is the communication flow inside the generated NCT. Additionally, only δ percentile of edges is presented in $w_{n,\delta}$. This allows for simulating the increase of new relationships among collaborators according to time. Each combination of w_c, $w_{n,\delta}$ is used as input for $fAnalyze$. All metrics are collected and stored in a result set RS. Once RS is complete, it is possible to plot charts, such as the ones presented in the experimental evaluation.

Note that Algorithm 4 takes as input the growth ratio δ ($0 \leq \delta \leq 100$). Initially, the edge weights for both G_c and G_n are randomly generated according to the same distribution. After that, Table 1 summarizes parameters adopted in experimental evaluation.

Algorithm 5 Network Growth

```
1: function NetGrowth(w_c, w_n, r)
2:     RS ← {}
3:     for all δ ← 0 to 100 step r do
4:         w_{n,δ} ← Filter(δ, δ/100 · w_n)
5:         w_{m,δ} ← fMix(w_c, w_{n,δ})
6:         RS ← RS ∪ fAnalyze(w_c, w_{m,δ})
7:     end for
8:     plotCharts(RS)
9: end function
```

5.3 Toy Sample Analysis

To better understand the mechanics of the growth ratio, we present a toy graph that corresponds to one of the smallest SME possible. It has ten vertices, two groups for G_c, and ten edges in both G_c and G_n ($v_c = v_n = 10$, $k_c = 2$, $e_c = e_n = 10$). Figures 3(a) and 3(b) are respectively examples of the CCT

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017

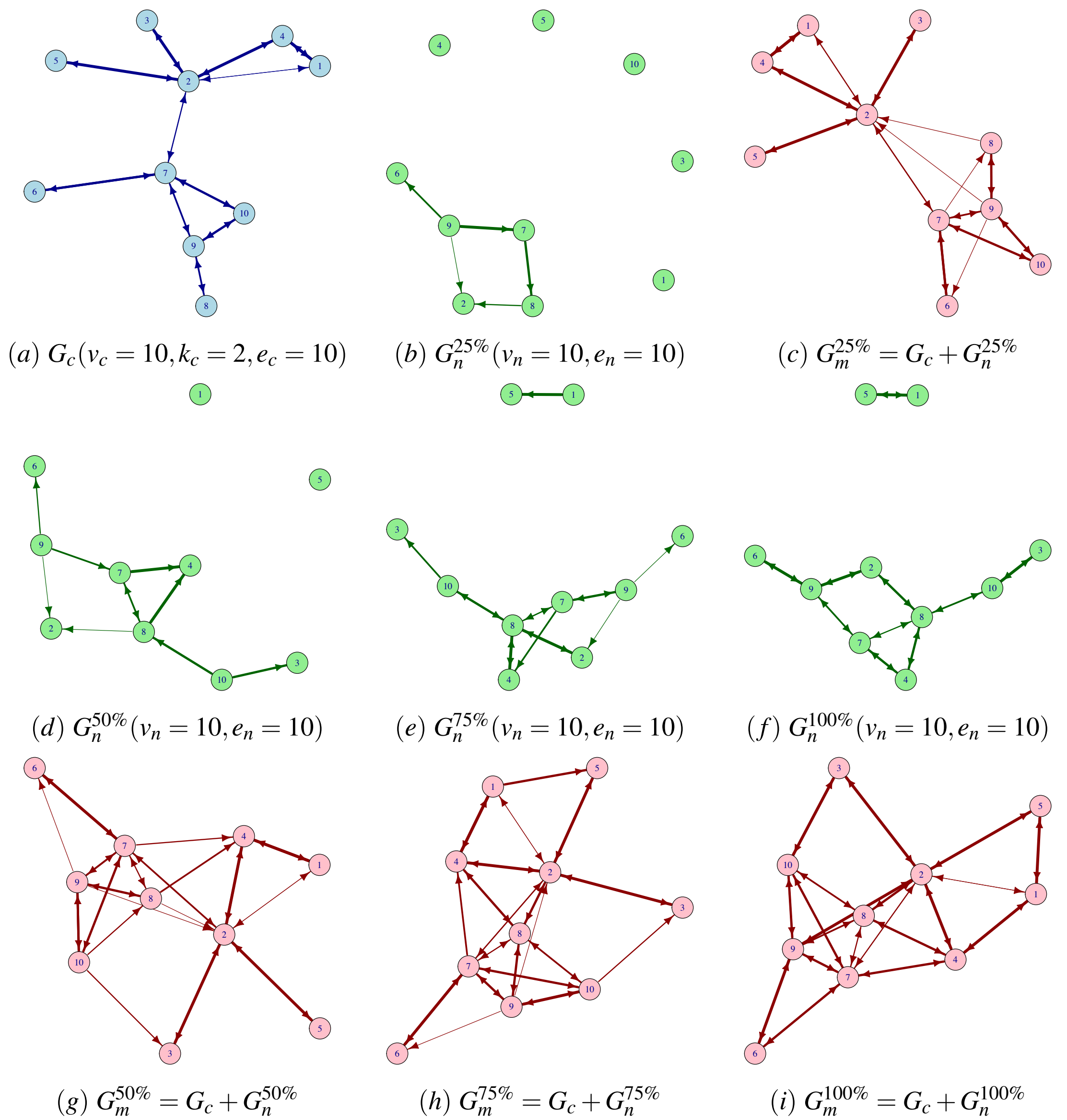


Figure 3. An example of current tool G_c (a) and new tool G_n (b) produced by Algorithm 4. The mixed graph G_m is produced by Algorithm 2 from both G_c and G_n. A network growth for new tool (G_n) with ration equals to 25% (b), 50% (d), 75% (e), and 100% (f); with their respectively effects in producing mixed graphs (G_m), $G_m^{25\%}$ (c), $G_m^{50\%}$ (g), $G_m^{75\%}$ (h), and $G_m^{100\%}$ (i). The width of edges are related to their weights

and the NCT graphs produced by Algorithm 4 according to this small setup. Figure 3(c) presents the produced mixed graph (G_m) from both G_c and G_n using Algorithm 2.

In the example, Figure 3(a) simulates communications that occurs through CCT inside the small enterprise. In this example, we assume that enterprise has two groups. It is possible to view some clusters of communication, which can be found among employees who share close relationship, such as work on related tasks, where the internal processes of the company make them to have a direct communication. Despite these clusters, it is possible to observe that the graph is connected. This means that with the mediation of one or more persons, the information can be disseminated through the network. In a small network like G_c, we can visually inspect the characteristics that are part of the goals of our analysis, such as connectivity, presence of clusters, and center points connecting them which are the employees

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017

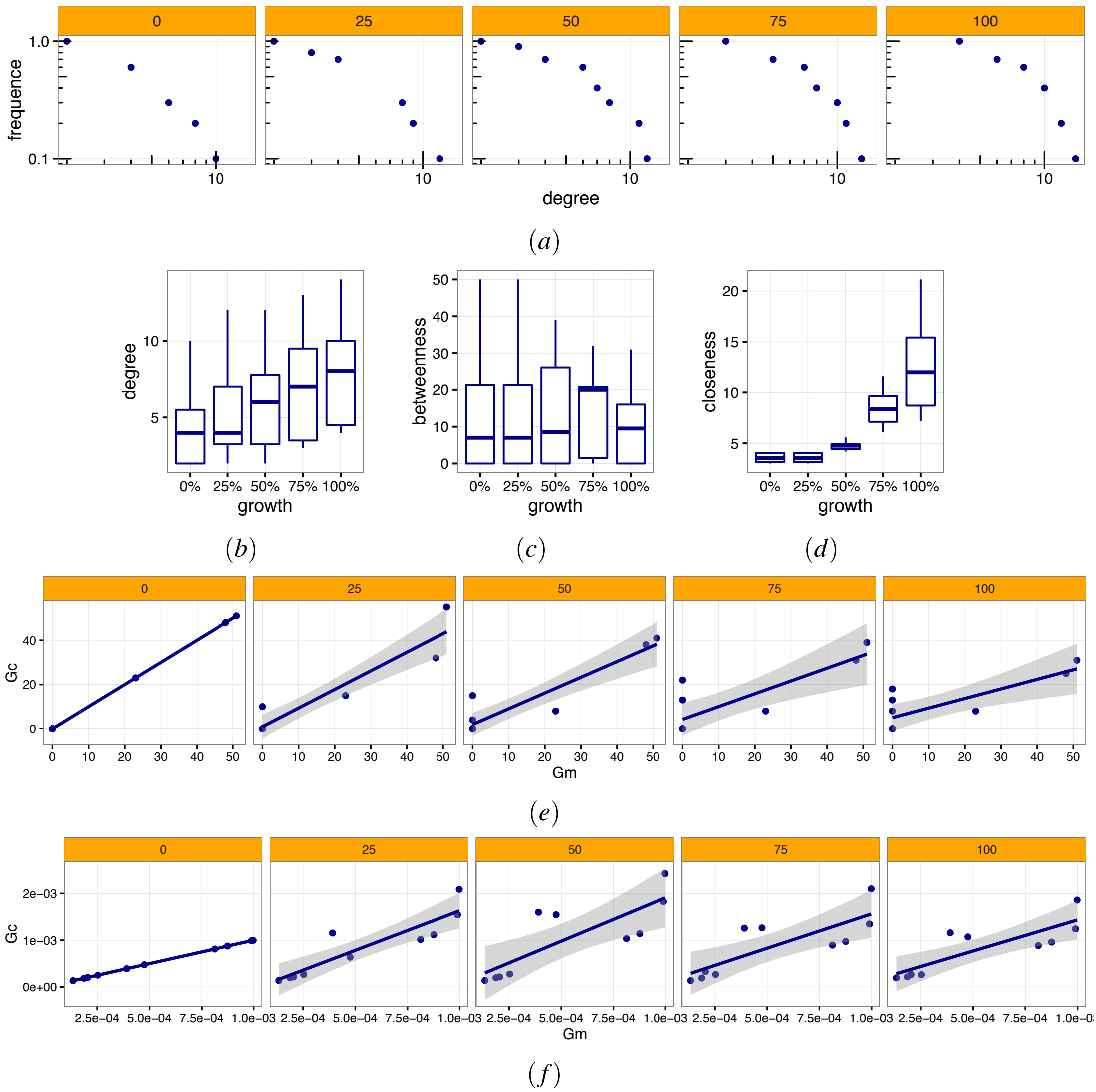


Figure 4. Descriptive statistics of G_m in the toy example grouped by growth ratio $\delta = \{0, 25, 50, 75, 100\}$. The degree distribution of G_m is in *log* x *log* scale (a). Box-plot of degree (b), closeness (c), and betweenness (d) distributions of G_m. Correlation plot of betweenness (G_c x G_m) (e). Correlation plot of closeness (G_c x G_m) (f)

identified as 2 and 7. Basically clusters communicate with each other through the central points. This is typically observed in a company where there are bridges of communication between areas or groups of people through coordinators, managers, and so on. We applied similar procedure to produce the graph associated to the NCT (G_n) depicted in Figure 3(b) and described in Algorithm 4.

Figure 3 explores different network growth (δ) of the new tool (G_n) using ratios such as 25%, 50%, 75%, and 100% in Algorithm 5. It is possible to observe that both the number of edges in G_n and their weights are explored in different growth ratios ($G_n^{25\%}$ (b), $G_n^{50\%}$ (d), $G_n^{75\%}$ (e), and $G_n^{100\%}$ (f)). This leads to different mixed graphs G_m: $G_m^{25\%}$ (c), $G_m^{50\%}$ (g), $G_m^{75\%}$ (h), and $G_m^{100\%}$ (i) by mixing G_c with G_n. By visually inspecting the instance of G_m presented in Figure 3, it seems that the communication flow is not restricted by the hierarchical structure as the growth ratio of G_n increases.

To better comprehend the toy sample, Figure 4 presents descriptive statistics for G_m produced by mixing $G_c(v_c = 10, k_c = 2, e_c = 10)$ with $G_n(v_n = 10, e_n = 10)$. Figure 4(a) depicts the frequency of degree of G_m as G_n grows. The degree of vertices increases as G_n grows. The plots in *log* x *log* scale fits a power law distributions, *i.e.*, suggesting a scale-free graph. This behavior is also summarized

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017

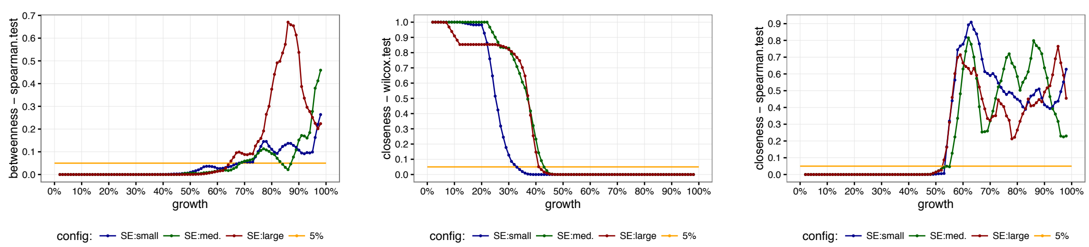
in Figure 4(b). Additionally, Figures 4(c) and 4(d) describe the closeness and betweenness centrality distribution. In Figure 4(d), the box plot for growth ratios of 50%, does not present any intersection with box plots of smaller growth ratios (0% and 25%). This indicates significant difference among them, *i.e.*, the median closeness of $G_m^{50\%}$ is higher than in G_c. Nevertheless, the betweenness described in Figure 4(c) does not present any significant difference among them.

Furthermore, Figures 4(e) and 4(f) present, respectively, a scatter plot for the closeness and betweenness correlation between G_c and G_m. The correlation is plotted with a confidence interval of 95%. It is possible to observe that both are correlated. This indicates, for example, that although Figure 4(c) indicates an increase in closeness introduced by G_n, such an increase does not change the topology of G_c, *i.e.*, it is not introducing a complementary behavior. It is actually just introducing an increase in scale of G_m with respect to G_c.

However, analyzing these plots may not be applicable in general, specially for larger networks. To tackle this problem, the MGF uses statistical analysis to assess and monitor the complementarity of NCT. It applies the Wilcoxon rank sum test and the Spearman rank correlation test to both betweenness and closeness as described in our Main Analysis.

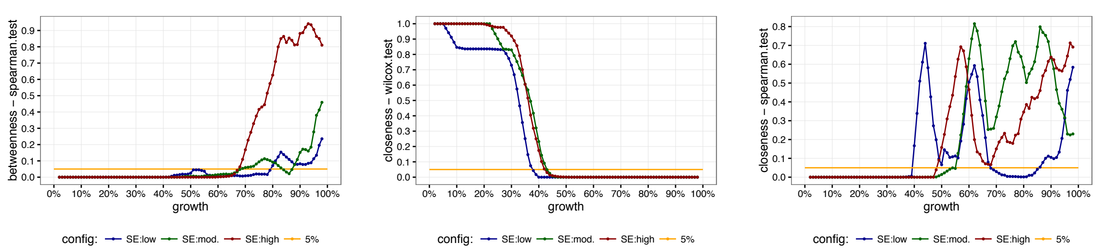
5.4 Sensitive Analysis

In this section, we evaluate the proposed MGF using synthetic data described in Section 5.1. It is worth mentioning that the objective of this section is not to evaluate the impacts of introducing a NCT. Instead, we intend to evaluate whether the MGF is able to distinguish G_c and G_m according to the influence of G_n. We have conducted a sensitivity analysis between networks. The goal is to identify if the NCT keeps the communication flow provided by CCT or if it introduces alternative and significant changes in the communication flows.



(a) (b) (c)

Figure 5. Scenario of Small Enterprise - varying number of edges in G_n: betweenness correlation analysis (a), closeness median analysis (b), closeness correlation analysis (c)

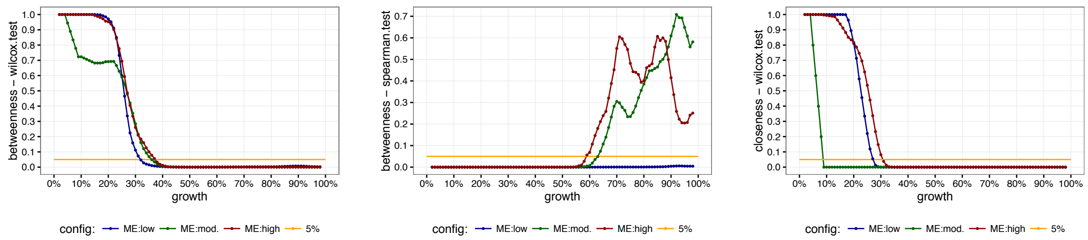


(a) (b) (c)

Figure 6. Scenario of Small Enterprise - varying number of groups in G_c: betweenness correlation analysis (a), closeness median analysis (b), closeness correlation analysis (c)

In the first scenario described in Table 2 we explored the number of communication flows in the NCT of SMEs under a small, medium, and large scale. In terms of betweenness, Figure 5(a) indicates a significant difference for the correlation, when the growth ratio is greater than 60%. Additionally, in terms

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017



(a) (b) (c)

Figure 7. Scenario of Medium Enterprise - varying both number of groups in G_c and number of edges in G_n: betweenness median analysis (a), betweenness correlation analysis (b), closeness median analysis (c)

of closeness, both median (Figure 5(b)) and correlation (Figure 5(c)) presents a significant difference when the growth ratio are greater than 40% and 55%, respectively.

We also explored a second scenario for SMEs, in which we vary the number of groups inside G_c. Figure 6(a) indicates a significant difference for the betweenness correlation when the growth increases. They were reached after a growth of 65%. In fact, the growth threshold for a significant difference occurs later when the group size is moderate or low. When it comes to closeness, both median analysis (Figure 6(b)) and correlation analysis (Figure 6(c)) present a significant differences when growth is greater than 40%. This is interesting as it indicates an increase in the number of messages in G_m and a difference in the network communication topology, as well.

In our third evaluation scenario, we explored the number of communication flows in the NCT and the number of groups inside G_c of a Medium Enterprise under small, medium, and large scale. In terms of betweenness, as depicted in Figures 7(a) and 7(b), we observe a significant difference for the median and correlation as the growth ratio reaches 35% and 60%, respectively. A similar behavior occurs with closeness. Figure 7(c) indicates a significant difference for the closeness median when the growth is greater than 35%. In fact, for the medium size case, only when reaching a growth greater than 70% we observe a clear significant difference. Before this value, we observed an oscillatory behavior around the significance threshold.

6 CONCLUSION

This paper proposes a Mixed Graph Framework (MGF), which aims at providing a set of quantitative approaches to analyze the complementary of a new communication tool (NCT) with relation to a current communication tool (CCT) in the enterprise context. This is done by measuring when the NCT brings significant differences in the overall enterprise communication flow with respect to the usage of the CCT. We model CCT and NCT communication interactions as the weighted graphs G_c and G_n, respectively. From these graphs, the MGF computes a mixed graph (G_m) that combines both G_c and G_n considering their usage. Our approach is then able to identify changes in overall communication within the enterprise.

We also evaluated the proposed MGF using synthetic data from which we have conducted a sensitive analysis. The sensitivity analysis is used to compare the weighted closeness and betweenness of both G_c and G_m. Our approach is able to identify whether G_n is providing any changes in the entire communication flow. It is worth mentioning that our approach does not propose adopting the NCT as a replacement for the CCT to promote communication empowerment. Instead, the goal of MGF is to aid managers in a decision-making process, giving them elements to conduct what-if analysis while deploying NCTs and measuring its influence in the entire set of communication solutions adopted in the enterprise.

We considered the evolution of a single network over time including time notion in the proposed framework as a promising future research. As well as performing case studies with networks of different sizes, which is a useful analysis for enterprises with scenarios of reorganizations, mergers and divisions of teams.

ACKNOWLEDGMENTS

The authors would like to thank CNPq, CAPES, and FAPERJ for partially funding this research.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017

REFERENCES

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Englewood Cliffs, New Jersey.

Barabási, A.-L. and Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512.

Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. (2009). Characterizing User Behavior in Online Social Networks. IMC '09, pages 49–62, New York, NY, USA. ACM.

Bennett, S. (2012). The benefits of communicating and collaborating in a real-time enterprise social network. pages 1–4.

Chai, S. and Kim, M. (2012). A socio-technical approach to knowledge contribution behavior: An empirical investigation of social networking sites users. *International Journal of Information Management*, 32(2):118–126.

Chen, H., Shen, H., Xiong, J., Tan, S., and Cheng, X. (2006). Social Network Structure Behind the Mailing Lists: ICT-IIIS at TREC 2006 Expert Finding Track.

Cross, R., Parker, A., Prusak, L., and Borgatti, S. P. (2001). Knowing what we know:: Supporting knowledge creation and sharing in social networks. *Organizational Dynamics*, 30(2):100–120.

Dalgaard, P. (2008). *Introductory Statistics with R*. Springer, New York, 2nd edition edition.

Devore, J. L. and Berk, K. N. (2011). *Modern Mathematical Statistics with Applications*. Springer, New York ; London, 2nd ed. 2012 edition edition.

Ebel, H., Mielsch, L.-I., and Bornholdt, S. (2002). Scale-free topology of e-mail networks. *Physical review E*, 66(3):035103.

Eurostat (2016). Statistics on small and medium-sized enterprises. Technical report, http://ec.europa.eu/eurostat/statistics-explained/index.php/Statistics_on_small_and_medium-sized_enterprises.

Friedman, B., Burns, M., and Cao, J. (2014). Enterprise social networking data analytics within Alcatel-Lucent. *Bell Labs Technical Journal*, 18(4):89–109.

Hamulic, I. and Bijedic, N. (2009). Social network analysis in virtual learning community at faculty of information technologies (fit), Mostar. *Procedia - Social and Behavioral Sciences*, 1(1):2269–2273.

Hansen, D., Shneiderman, B., and Smith, M. A. (2010). *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. Morgan Kaufmann, Amsterdam, 1 edition edition.

Hoffmann, W. H. and Schlosser, R. (2001). Success Factors of Strategic Alliances in Small and Medium-sized Enterprises—An Empirical Survey. *Long Range Planning*, 34(3):357 – 381.

Johnson, R., Kovács, B., and Vicsek, A. (2012). A comparison of email networks and off-line social networks: A study of a medium-sized bank. *Social Networks*, 34(4):462–469.

Kim, J., Leem, C., Kim, B., and Cheon, Y. (2013). Evolution of Online Social Networks: A Conceptual Framework. *Asian Social Science*, 9(4):208.

Kleinberg, J. M. (1998). Authoritative Sources in a Hyperlinked Environment. SODA '98, pages 668–677, Philadelphia, PA, USA.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer, 2009 edition edition.

Larsen, R. J. and Marx, M. L. (2005). *An Introduction to Mathematical Statistics and Its Applications*. Prentice Hall, Upper Saddle River, N.J, 4 edition edition.

Layman, L., Williams, L., Damian, D., and Bures, H. (2006). Essential communication practices for Extreme Programming in a global software development team. *Information and Software Technology*, 48(9):781 – 794.

Leftheriotis, I. b. and Giannakos, M. (2014). Using social media for work: Losing your time or improving your work? *Computers in Human Behavior*, 31(1):134–142.

Lynn, T. G., Healy, P., Hunt, G., and Morrison, J. P. (2015). Towards a General Research Framework for Social Media Research Using Big Data. IEEE.

Nettleton, D. F. (2013). Data mining of social networks represented as graphs. *Computer Science Review*, 7:1–34.

Newman, M. (2003). The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256.

Newman, M. E. J., Watts, D. J., and Strogatz, S. H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572.

Ngai, E., Tao, S., and Moon, K. (2015). Social media research: Theories, constructs, and conceptual

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017

frameworks. *International Journal of Information Management*, 35(1):33–44.
Ning, K., Li, N., and Zhang, L.-J. (2012). Using Graph Analysis Approach to Support Question & Answer on Enterprise Social Network. volume 0, pages 146–153, Los Alamitos, CA, USA. IEEE Computer Society.
Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*, 32(3):245–251.
Prado, A. B. and Baranauskas, M. C. C. (2013). Addressing structural and dynamic features of scientific social networks through the lens of Actor-Network Theory. *Social Network Analysis and Mining*, 3(4):1263–1276.
Raghavan, P. (2002). Social networks: from the Web to the enterprise. *IEEE Internet Computing*, 6(1):91–94.
Rooksby, J. and Sommerville, I. (2011). The Management and Use of Social Network Sites in a Government Department. *Computer Supported Cooperative Work (CSCW)*, 21(4-5):397–415.
Stewart, S. A. and Abidi, S. S. R. (2012). Applying social network analysis to understand the knowledge sharing behaviour of practitioners in a clinical online discussion forum. *Journal of Medical Internet Research*, 14(6):e170.
Turban, E., Bolloju, N., and Liang, T.-P. (2011). Enterprise social networking: Opportunities, adoption, and risk mitigation. *Journal of Organizational Computing and Electronic Commerce*, 21(3):202–220.
Wang, Y., Iliofotou, M., Faloutsos, M., and Wu, B. (2011). Analyzing interaction communication networks in enterprises and identifying hierarchies. pages 17–24.
Wasko, M. M., Teigland, R., and Faraj, S. (2009). The provision of online public goods: Examining social structure in an electronic network of practice. *Decision Support Systems*, 47(3):254–265.
Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, 1 edition edition.
Zaffar, F. O. and Ghazawneh, A. (2012). Knowledge sharing and collaboration through social media–the case of IBM.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3114v1 | CC BY 4.0 Open Access | rec: 28 Jul 2017, publ: 28 Jul 2017