Mutual information to assess structural properties in dynamic networks

David Rodrigues and Jorge Louçã

ISCTE – Lisbon University Institute

Av. Forças Armadas, 1649-026 Lisbon, Portugal

E-mail: David.Rodrigues@gmail.com, Jorge.L@iscte.pt

Abstract. This article proposes applying the *variation of information* measure from Information Theory to evaluate macro-level properties characterising dynamic networks. This measure is used to evaluate different clusters given by the agglomerative hierarchical clustering algorithm of Clauset, Newman and Moore (2004), concerning a case study of the multi-agent based network model of a university email service. The *variation of information* measure is shown to be capable of assessing the outcome of simulating the dynamics of networks, in terms of its macro-level properties.

Keywords: social networks, community detection, clustering algorithms, multi-agent simulation, information theory, email

1. Introduction

Informal communication networks, arising from the interactions of a multitude of individuals, have a role in solving problems and in sharing knowledge between members of organisations (Tyler, Wilkinson, & Huberman, 2003). Informal communication networks are frequently characterised by some sort of auto-organisational phenomena. Typically appearing social structures are communities or modules, defined as having a higher density of connections between members than the density of inter-module connections (Girvan & Newman, 2001). The study of these community dynamics, including analysis, characterisation of main features, and prediction of their evolution, is becoming more and more relevant for the understanding of human social behaviour.

The study of community dynamics implies some methodological options. One of them concerns what to observe when studying network dynamics. An idea is to apply community detection algorithms to data sets representing different moments in time. This allows depicting communities' evolution from real data sets. We argue that, in this context, simulation models can be designed and used to predict the evolution of community configurations. An important issue is, then, how to evaluate the quality of the simulation results. We propose to use metrics from Information Theory, namely the *variation of information* measure (Meilă, 2007), allowing to calibrate a simulation model from real data, and to assess the macro-level properties of the resulting communities.

This article starts by reviewing the most relevant community detection algorithms, focusing on hierarchical algorithms based on modularity. Particularly, some previous work on the realm of email community detection is referenced. Then, the case study of a university email network is depicted. The email communication network of the Lisbon

University Institute is transversal to the hierarchical structure of the organisation, and characterised by the existence of diverse informal communities. Results from this modelling are used to test how the metric of *variation of information* can be applied to clustering analysis. Finally, a multi-agent simulation based on the email network is presented. The *variation of information* measure is then used to assess the clustering property on the simulation results vs. the observed email network dynamics.

2. State of the art

This section presents the theoretical concepts used in this research, namely: (1) the domain of community detection algorithms, with a particular focus on hierarchical clustering techniques, (2) the research found in the literature concerning applying community detection to email networks, and (3) the *variation of information* measure, from Information Theory, that we apply to the measurement of distance between clusters.

2.1 Community detection algorithms

Community detection can be very useful in performing an exploratory analysis of data, and its usage transverses several domains, from statistics to computer science, biology to psychology. In every scientific domain it is necessary to deal with empirical data. One of the first classifications that one tries is to group the data according to some property that might manifest itself similarly inside the groups. Several algorithms and techniques have been devised to accomplish this partitioning, allowing for a variety of solutions. Some methods are robust, and can be used effectively to classify groups within heterogeneous sets of data. On the other hand, some algorithms are specific to certain problems, and require particular initial conditions (Shortreed, 2006).

Network clustering techniques can be divided into global or local classes of algorithms. Local algorithms use some local pattern to determine which points belong to each cluster. For instance, clique percolation (Palla, Derényi, Farkas, & Vicsek, 2005) uses the notion of clique to identify groups or modules. In global strategies, the network is taken as a whole and a general property is used to divide the network, separating all its members into clusters. One example is hierarchical clustering techniques, like the Girvan-Newman algorithm (Girvan & Newman, 2001), using edge betweenness as the property of interest. These techniques are discussed in the following paragraphs.

Hierarchical clustering is a type of partitioning strategy that produces a dendrogram from the breaking down of a complete graph. Two subclasses of this type of partitioning are available, according to the way the dendrogram is built. One concerns a bottom-up approach, where firstly each node is considered a member of its own community, and then the process runs iteratively, merging communities according to some maximal value of a quality function. These are called hierarchical agglomerative methods. Another alternative is the divisive methods of hierarchical clustering, where all nodes belong to one single initial community. The dendrogram is designed by breaking the communities iteratively into sub-communities, up to the point where all nodes are attributed to communities with one node. The way divisions are made is based on the 'optimal' value of some property, usually one that measures the strength of connections between communities. This subclass of hierarchical clustering methods is known as hierarchical divisive.

Two examples of hierarchical clustering methods are the Girvan-Newman algorithm (GNA) (Girvan & Newman, 2001), and the Clauset-Newman-Moore algorithm (CNM) (Clauset et al., 2004). The former is a hierarchical divisive algorithm, while the latter is agglomerative. GNA uses the edge betweenness to determine which edges can be safely removed from the network and iteratively removes them, splitting the network into sub-networks to construct the final dendrogram. The point in the dendrogram where the 'optimal' cut is achieved is then given by a property called modularity (Newman & Girvan, 2003). This property is based on the notion of assortative mixing, describing the fact that some networks present a tendency for connecting vertices to other vertices that have some sort of similarity (Newman, 2002). In CNM, the modularity itself is used as the measure to determine which sub-communities to join, and from all different joining possibilities, it chooses the one which presents the maximal value of modularity. The 'optimal' point at which to cut the dendrogram is then where the value of modularity is maximal (Clauset et al., 2004). Both of these algorithms use the notion of modularity to determine the optimal point at which to cut the dendrogram (Newman, Barabasi, & Watts, 2006; Newman, 2006).

2.2 Community detection in email networks

The community detection algorithms presented above have been designed to analyse different types of data. One of these types concerns informal communication networks, namely email networks. Research found in the literature describes the use of community detection algorithms, aiming to distinguish specific kinds of messages, like spam, and more generally to identify the informal structure of email networks.

One of the greatest interests in email classification approaches is due to the necessity for filtering email to avoid the spread of spam. Several methods for the detection and elimination of spam messages have been proposed: tests based on semantic analysis, Bayesian tests, black lists and white lists. These methods can give good results, but created another problem: false positives (Garriss et al., 2006). A different approach, recently employed, takes advantage of the knowledge that one might have of the social structure that the user is in. This technique, proposed by Kim (2007), uses a spectral decomposition of the Laplacian matrix, built from the headers of the messages received in the user's inbox. From this decomposition of the network of contacts into sub-networks, each of the sub-networks is classified according to its clustering coefficient. Kim states that sub-networks where the clustering coefficient is high can be classified as spam-free. In his tests, the author achieved 100% accuracy in the classification of emails from a test data set. Besides the natural interest in detecting and classifying spam email, the spectral decomposition analysis can be directed to legitimate email traffic, in order to find other kinds of properties that emerge from the exchange of messages.

In a study made at the Hewlett-Packard Labs, Tyler et al. (2003) conducted an analysis of the email logs to detect the informal structure of the relations and interests networks that were formed. With this study the authors tried to understand the dynamics of information inside the company laboratory. Informal networks were shown to coexist along with the formal organisational structure, serving the organisation at different levels, like conflict resolution or the definition of internal projects. It was also found that informal communication networks work as a means of learning and knowledge transmission inside the company. Due to the value that these communities present for organisations, an automated method was developed that identifies sub-networks. A set of email logs was used, including one million exchanged messages over a period of two months. A minimum

of messages between two people was defined in order for the existence of a connection to be accounted for. This minimum number of messages between any two vertices defines the threshold that one can vary to construct the graph. The graphs constructed in this manner were found to have a power-law for high thresholds. The algorithm used to divide the graph into modules was based on Freeman's betweenness (1977) applied to the edges. This is similar to the Girvan-Newman algorithm (Girvan & Newman, 2001).

Also concerning the detection of informal structures, Ebel et al. (Ebel, Mielsch, & Bornholdt, 2002), at the University of Kiel, studied an email network from the logs of email servers for a 112-day span. Each node of the networks corresponds to a student, and the connections put in evidence the passing of messages between them, with an average degree of 2.88. Also, it was verified that the degree distribution falls under a power law with an exponential behaviour in the tail of the distribution (for degree>100). The power law had an exponent of 1.81. When measuring the clustering of the network, the authors found that the network had a neighbourhood effect due to the emails exchanged with the outside of the network, lowering the clustering values, but still one or two orders of magnitude greater than the expected value for random networks with equal degree distribution.

In this work, hierarchical clustering methods were used to create a clustering from the graphs that result from the simulation model, and then this clustering was used to calculate the variation of information from the real known clustering. The variation of information method is described in the following point.

2.3 Variation of information as a measure of clustering distances

The need to assess how appropriate the result of a partitioning procedure is leads to the idea that some sort of measure is needed. This measure should be capable of giving information when applied both to real data and to simulation results. Information based measures are appropriate if one wants to find how similar two different clusters are. In particular, aiming to empirically assess the quality of a clustering algorithm by comparing it to a know reality, one needs to define a distance or a measure on the space of the data set (Meilă, 2007).

We propose to use the measure called *variation of information* (VI), which allows measuring the amount of information lost and gained in changing from clustering C to clustering C' of the same data set (Meilă, 2007). The following paragraphs present the *variation of information* measure, synthesising a more detailed description by Meilă (2007).

Considering one partitioning C, the probability that a node k belongs to a cluster C_k is given by the equation (1) where n_k is the number of nodes in the cluster C_k and n is the total number of its elements.

$$P(k) = \frac{n_k}{n} \qquad (1)$$

The uncertainty associated with this measure is given by the entropy of the variable $P(k)$

$$H(C) = -\sum_{k=1}^{K} P(k)\log(P(k)) \qquad (2)$$

where $H(C)$ is the entropy associated with the clustering C. This is always a non-negative value, being zero when there's no uncertainty. The mutual information between two clusterings, C_k and C'_k, represented by $I(C_k, C'_k)$, means the information that one has over the other. The mutual information is given by the probability $P(k,k')$, representing the probability that a node belonging to the cluster C_k is in the cluster $C'_{k'}$.

$$P(k,k') = \frac{\left|C_k \cap C'_{k'}\right|}{n} \tag{3}$$

Using expression (3), the mutual information $I(C,C')$ is defined as the mutual information associated with the two random variables k and k':

$$I(C,C') = \sum_{k=1}^{K}\sum_{k'=1}^{K'} P(k,k')\log\frac{P(k,k')}{P(k)P(k')} \tag{4}$$

Meilă (2007) proposed *variation of information* (VI) as a criterion to compare both clusters:

$$VI(C,C') = H(C) + H(C') - 2I(C,C') \tag{5}$$

This measure is a metric; since it is always non-negative, it is symmetric and presents triangular inequality. The *variation of information* can be normalised, taking into account the conditions of the clusters. If the normalisation concerns the same data set, we obtain

$$V(C,C') = \frac{1}{\log n} VI(C,C') \tag{6}$$

where n is the number of nodes. If the normalisation does not concern the same data set, but still has the same number of clusters in both clusters, it is possible to normalise the *variation of information* via

$$V_{k*}(C,C') = \frac{1}{2\log K*} VI(C,C') \tag{7}$$

where $K*$ is the number of clusters in each cluster.

The *variation of information* criterion for comparing two clusterings of a data set is derived from information theoretic principles. VI makes no assumption about how the clusterings may be generated, and requires no rescaling to compare values of the two clusterings C and C'. Also, VI doesn't directly depend on the size of the data set. These qualities allow using the VI measure when comparing across different data sets and different clustering algorithms.

In this review of the state of the art, we show some of the principles and methods used in community detection, namely those used in email networks, and present one information-based measure for the comparison of partitioning methods. The following section presents the application of this measure to a case study, assessing results of the partitioning obtained by the community detection algorithms above. Then, a multi-agent model for simulating the communication network composition process is presented. The

variation of information measure is then used to assess the behaviour of the simulation via the partitioning.

3. Modelling structural dynamics in email networks

Our proposal aims to show how *variation of information* can be used to assess the results of multi-agent simulations. For this, we developed a multi-agent model of the university email system, to represent the process of forming a communication network between professors. Real data was collected from the logs of email servers for a period of 62 days. From this data the logs where anonymised under a privacy directive of the institution, and then the teacher network was drawn up. The teacher network was considered undirected and was composed of 395 nodes with a density of 2.69%. This network was then analysed through the community detection algorithm proposed by Clauset, Newman and Moore. The model includes the possibility of using real data as training. We also assess the influence of the *social neighbourhood* of the nodes in the structure of the network.

3.1 Results from detecting communities in the communication network

Two global hierarchical clustering algorithms were used to analyse the professors' communication network: the Girvan-Newman and the Clauset-Newman-Moore algorithms. Both use modularity as a quality function for the partitioning. Investigation of the overlapping communities was done via the clique percolation algorithm, and a k-core analysis was also carried out to find hierarchical clusters inside the teachers' network. The results are summarised in Table 1.

Table 1 – Analysis of the teacher network for different clustering methods

Method	*Groups*	*Modularity (Q)*
Girvan-Newman	14	0.588
Clauset-Newman-Moore	7	0.585

The number of departments/sections in the university log files is 14. The Girvan-Newman algorithm found the same number of components. However, the Clauset-Newman-Moore gave only seven communities. Both of the algorithms had a maximal value of modularity (Q) near 0.6 (0.588 for the former and 0.585 for the latter) and presented communities that didn't match exactly those defined by the departments of the university, indicating a transversal communication process.

Both algorithms were compared with the real clustering of the university departments. The results for the *variation of information* and the *equivalent random clustering value* are shown in Table 2. The *equivalent random clustering* is obtained by keeping the same number of clusters and maximising the uncertainty by equally distributing its members.

Table 2 – Comparison of Variation of Information of two clustering algorithms with respect to the real departments of the Lisbon University Institute

Clustering	*Variation of Information (V)*
Girvan-Newman	0.224
Equivalent random clustering	0.833
Clauset-Newman-Moore	0.236
Equivalent random clustering	0.718

Table 2 shows that both algorithms are approximately at the same distance from the departmental clustering of the university. Also, the low values for the normalised *variation of information* (V), when compared to equivalent random clustering, reveal that the clustering algorithms produced clusters very close to the real departments found at the university. Differences are due to cross-department communication, and this has been verified in other research (Rodrigues, 2009) via the clique percolation method (Palla et al., 2005).

3.2 Training mode vs. free regime

We designed a multi-agent that captures the informal communication network formation between professors at the Lisbon University Institute. The model is founded on a real data set, and intends to predict the structural dynamics of the network. Initially the model is populated with the number of professors equal to those in the training set. Each of these professors will have an empty contact list, and during the subsequent phases they are trained with real data, reproducing the behaviours of the training set. The notion of time is given by a sequence of events. At each step only one event, for instance the sending of one email message, is reproduced.

At the end of each time step, the number of previous contacts of each professor is updated. Also, a probability of contacting any of the agent contacts is updated according to the previous contact's cardinality. This process goes on during the training phase of the model. When this phase is exhausted, the model starts evolving in a free regime, where at each time-step one agent will be given the opportunity to produce an event. Events are then governed by the principles of *high transitivity*, *assortative mixing* and *random events*, described as follows:

High transitivity

A high value of clustering indicates the existence of a high number of triangles in the network. The formation of a high transitivity network was accomplished through the inclusion of the idea of *social neighbourhood*. This distance allows including all the agents falling under a certain geodesic distance in a list of professors to contact when a certain event occurs. For example, consider three agents: *ego*, *alter1* and *alter2*. *alter2* is at a distance *d* from *ego* and 1 from *alter1*. A simple rule to account for the probability of establishing a contact between *ego* and *alter2* is defined as:

$$p^*_{ego,alter2} = \frac{1}{d^a} p_{alter1,alter2} \qquad (8)$$

The value of p^* represents an adjustment of the probability of connection over a certain distance. The parameter a is a measure of how distance influences this probability. To calculate this probability over all possible geodesic paths, the probabilities of connection assuring that $\sum p^*_{i,l} = 1$ are normalised, giving the normalised probability $p_{i,l}$:

$$p_{i,l} = \frac{\sum_{GeoPath} p^*_{i,l}}{\sum_k \sum_{GeoPath} p^*_{i,l}} \qquad (9)$$

Assortative mixing

The idea that agents prefer to connect other similar agents is achieved through the value of their degree. For a set of agents X=<x_1, ..., x_n> with a degree distribution <k_1,... ,k_n> and a historical probability distribution <p_1,...,p_n>, the corrected probability *pc* for the connection between agents *i* and *j* is given by:

$$pc_{i,j} = \frac{\frac{p_j}{1+|k_i - k_j|}}{\sum_j \frac{p_j}{1+|k_i - k_j|}} \qquad (10)$$

Random Events

The coupling of the sub-models through transitivity and assortative mixing isn't enough to guarantee obtaining a large network component. This happens because both sub-models only allow connections between nodes that are already connected (degree >0). This implies that new connections won't be created to isolated nodes. It is then necessary to introduce a residual probability of, for each event generated, the possibility of connection to/from someone with degree zero.

These three principles of *high transitivity*, *assortative mixing* and *random events* comprise the rules by which each agent acts when the simulation is placed under a free regime. This means that, when agents are given the opportunity to generate an event, they act according to these principles. First they define a set of other agents that they might contact, based on the *social neighbourhood*. The probabilities of contact are established according to the transitivity principle, and the historical number of contacts. Then these probabilities are corrected, taking into account the assortative mixing principle. From the final corrected probabilities an event is generated. Finally, agents apply the random event principle for the residual growth of the network. This is done when the simulation is in free regime. Otherwise, in training regime the agents limit themselves to reproducing events of the training set, and update the historical number of contacts accordingly.

The outcome of this model is presented in the next section, where the training set was defined as a fraction of the total data and the model was tested for different *social neighbourhoods*.

4. Results and analysis

The model was run to test the influence of *social neighbourhood* and the influence of the training set in the final results. The resulting network degree, average path length, and clustering coefficient were measured. The Clauset-Newman-Moore algorithm was applied and the resulting clusters compared to the real departments, through the *variation of information* process. The results are summarised in this section.

The model was initially tuned for the value of the residual probability, with 50% of training data, neighbourhood 1 and the parameter $a=4$. The control value chosen was the average degree, which had to be short of 10% of the real data. We found a value of 4.0×10^{-4} for the residual probability of events to be adequate.

The real data showed that the average degree grows according to a power law of 0.593. For every run we calculated the average degree, the density of the network, the average path length and the clustering coefficient. For all of these measures the Clauset-Newman algorithm was run to indicate the value of maximal modularity of the network.

The evolution of the performance of the simulation, according to the training fraction of the real data, has been measured through the average degree (Figure 1), the average path (Figure 2), the clustering coefficient (Figure 3), the modularity (Figure 4), and finally the *variation of information* (Figure 5).

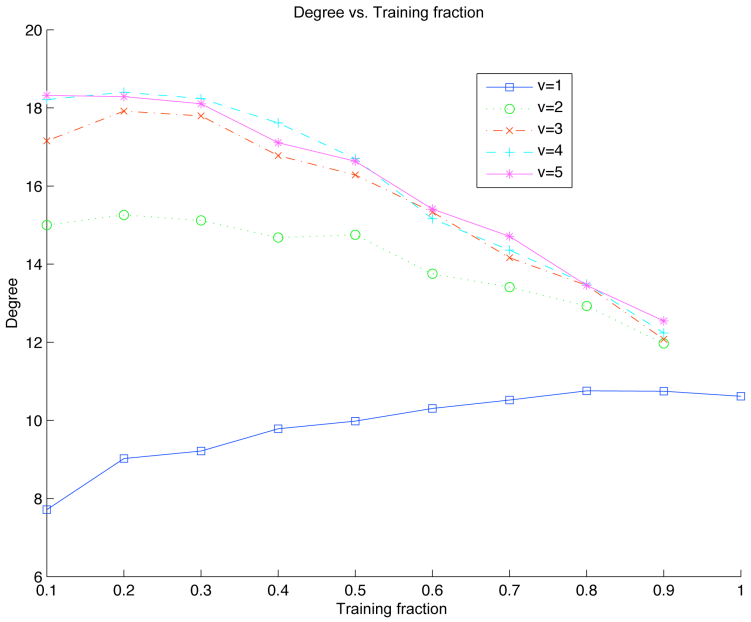

Figure 1 – Average degree vs. training fraction and neighbourhood (v)

The results represented in Figure 1 show that the final average degree of the simulation, for the case of low training fraction coupled with a neighbourhood (*v*) greater than 1, is very different from the observed real data. The training fraction equals 1 when the model uses exclusively real data. Also, the increase in neighbourhood seems to tend to a limiting impact.

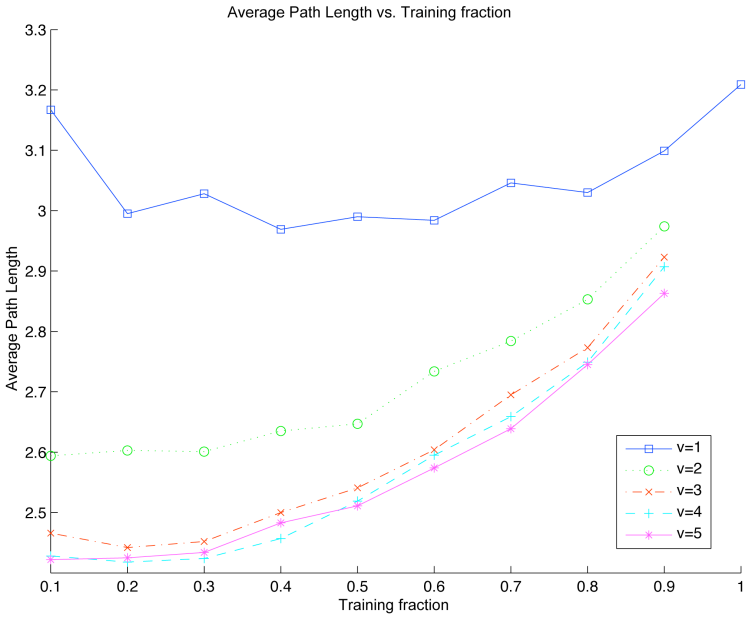

Figure 2 – Average path length vs. training fraction and neighbourhood (v)

The training fraction seems not to have a strong influence on the final average path length of the network, for a neighbourhood of 1, as the final results are almost constant. On the other hand, the usage of a value greater than 1 for the neighbourhood has a strong impact on the average degree, presenting simulation results that are lower than expected. There is a limiting factor where, for high values of *v,* the results tend to a limiting plateau.

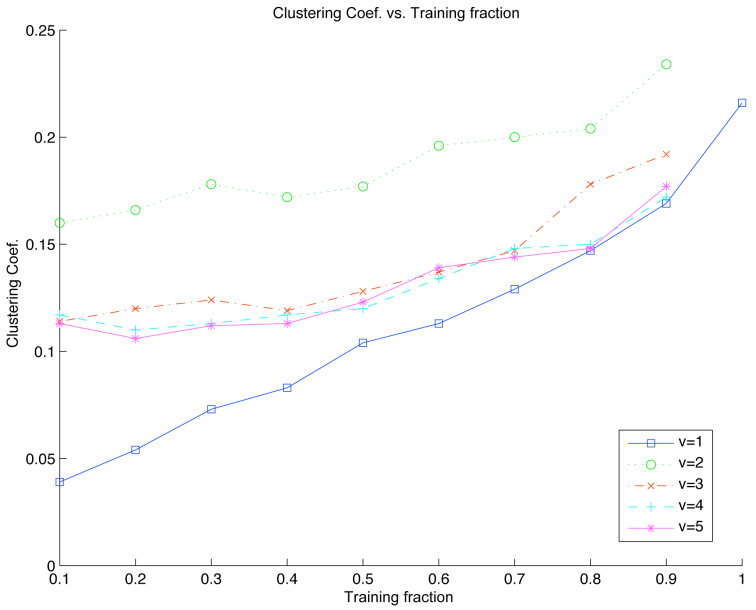

Figure 3 – Clustering coefficient vs. training fraction and neighbourhood (v)

Measuring the clustering coefficient showed that the model with neighbourhood 1 presents low values for low training fractions. This was expected, as it corresponds to a situation where connections are only possible with previously connected agents. When the value of the *social neighbourhood* is increased (v=2), the results show that even with small training fractions the model presents clustering values similar to those of the real data. Curiously, an increase in the value of the *social neighbourhood* has a negative effect on the coefficient clustering. This seems to indicate that first-order neighbours and their *alters* are the most important agents for the establishment of informal social networks.

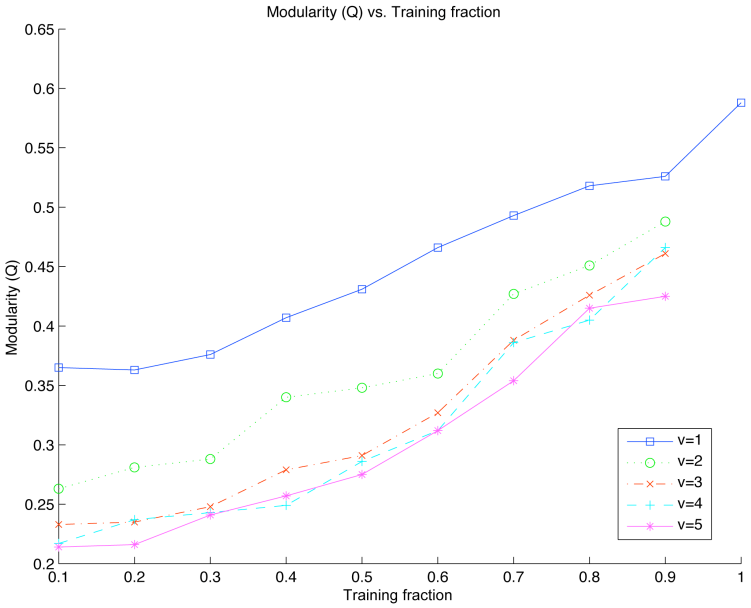

Figure 4 – Modularity vs. training fraction and neighbourhood (v)

Concerning the structure of the formed communities, it is remarkable that an increase in the neighbourhood has a negative effect on the value of modularity. If one thinks of communities as modules with higher edge density than the density of inter-community edges, the increase of the *social neighbourhood* will allow border nodes to have a higher probability of connecting agents from other communities, and therefore increase the inter-community edge density. In terms of the training fraction, even for 10% of training, the simulation still presents a structure (Q>0.3) for v=1.

Figure 5 shows the normalised *variation of information*, plotted as a function of the training fraction and for different *social neighbourhoods* (*v*). The horizontal lines represent the values of the *variation of information* between clusters. The bottom line represents the variation of information between the clustering produced by the Clauset-Newman-Moore algorithm and the real departments of the university, while the top line represents the *variation of information* of the clusters of a random network, with the same department distribution and the same number of clusters as the real network. The former represents the minimal distance, while the latter represents the highest distance between two clusters. Figure 5 shows a very low influence of *social neighbourhood* in terms of the final *variation of information*, although the effect is more important for higher training fractions. The *variation of information* decays linearly with the increase of the training fraction.

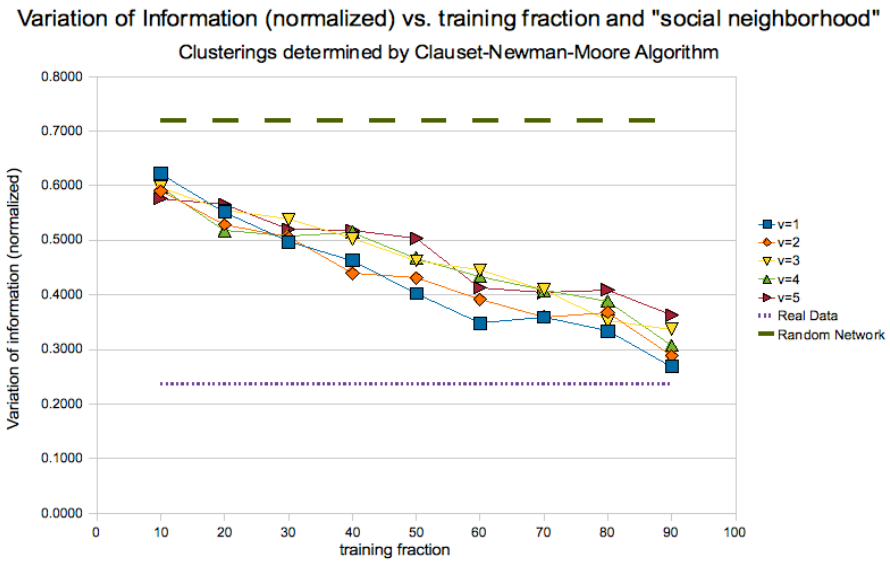

Figure 5 - Variation of Information for v=1...5 and training fraction 10%...90%

The modelling of the communication network formation showed that *social neighbourhood* has a strong influence on the average path length. This influence is remarkable in Figure 2, where for neighbourhood (v), greater than 1, the value of the average path length significantly diminishes. On the clustering coefficient side, it seems that the increase in neighbourhood for v>2 has a counter effect on clustering, as the number of possible *alters* to whom the *ego* can connect is higher, not producing immediate triangles as seen in Figure 3. On the other hand, Figure 4 shows that even for small training fractions the resulting network keeps its structure. The effect of higher neighbourhood destroys the boundaries of communities, as it increases the possibility that border members establish connections to members of other communities.

The *variation of information* measurement of the results from the simulation is shown in Figure 5, showing dependence on the training fraction that is used by the simulation. The *variation of information* is maximal between the clusterings obtained from low training fractions vs. the real departments. Also, the *variation of information* is minimal for a small *social neighbourhood*. This is consistent with the observed behaviour in modularity.

5. Conclusion and perspectives

This work shows the application of an information-based measure for quantifying dynamics in communication networks. The clustering produced by the application of a hierarchical algorithm is a macro-level structure. The dynamics of this structure were analysed in terms of *variation of information*. A multi-agent based simulation model was designed to represent the informal communication network in a university. The model used real data to train and adjust its parameters to the university population dynamics. The multi-agent modelling revealed that, even for

small fractions of training data, the communication network still presented significant structure, mainly for a *social neighbourhood* of one. The effect of this *social neighbourhood* is more important for low training fractions and as a positive effect in the value of the clustering coefficient for a small neighbourhood greater than one. In fact, the establishment of the informal communication processes inside the university presents some transitivity effect, from contacts of *ego alters,* but the real value of this effect is still not totally clear. We conclude that there is at least a qualitative influence of the *social neighbourhood* in the resulting network.

The use of the *variation of information* measure allows evaluating the distance between the results of a multi-agent simulation and known real data. Figure 5 shows the clear dependence of *variation of information* upon different *social neighbourhoods* and different training fractions.

In future research, the use of information-based measures will be extended to other kinds of multi-agent based simulations, aiming to prove their generality.

REFERENCES

Clauset, A., Newman, M.E.J., & Moore, C. (2004, August). Finding community structure in very large networks. *cond-mat/0408187*, from http://arxiv.org/abs/cond-mat/0408187 http://dx.doi.org/10.1103/PhysRevE.70.066111

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

Freeman, L.C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry, 40*(1), 35-41.

Garriss, S., Kaminsky, M., Freedman, M.J., Karp, B., Mazières, D., & Yu, H. (2006, May). *Abstract RE: Reliable Email.* Paper presented at the Proceedings of the 3rd Symposium on Networked Systems Design and Implementation, San Jose, CA.

Girvan, M., & Newman, M.E.J. (2001, December). Community structure in social and biological networks. *cond-mat/0112110*, from http://arxiv.org/abs/cond-mat/0112110

Kim, U. (2007). Analysis of Personal Email Networks using Spectral Decomposition. *International Journal of Computer Science and Network Security, 7*(4), 185-188.

Meilă, M. (2007). Comparing clusterings—an information based distance. *J. Multivar. Anal., 98*(5), 873-895.

Newman, M., Barabasi, A.-L., & Watts, D.J. (2006). *The Structure and Dynamics of Networks:* (1 ed.). Princeton, NJ: Princeton University Press.

Newman, M.E.J. (2002, September). Mixing patterns in networks, *cond-mat/0209450*, from http://arxiv.org/abs/cond-mat/0209450

Newman, M.E.J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences, 103*(23), 8577-8582.

Newman, M.E.J., & Girvan, M. (2003, August). Finding and evaluating community structure in networks*. cond-mat/0308217*, from http://arxiv.org/abs/cond-mat/0308217

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature, 435*, 814-818.

Rodrigues, D. (2009). *Detecção de Comunidades no Sistema de Correio Electrónico Universitário.* ISCTE—Instituto Superior de Ciências do Trabalho e da Empresa, Lisboa.

Shortreed, S. (2006). *Learning in Spectral Clustering.*

Tyler, J.R., Wilkinson, D.M., & Huberman, B.A. (2003). Email as spectroscopy: automated discovery of community structure within organisations (pp. 81-96): Kluwer, B.V.