Proceedings of the 9th International Conference on Applied Informatics
Eger, Hungary, January 29–February 1, 2014. Vol. 1. pp. 113–120
doi: 10.14794/ICAI.9.2014.1.113

The community structure of word association graphs

András Bóta[a], László Kovács[b]

[a]University of Szeged, Institute of Informatics
bandras@inf.u-szeged.hu

[b]University of West Hungary, Department of Applied Linguistics
klaszlo@btk.nyme.hu

Abstract

One of the defining characteristics of small-world networks is that their edge distribution is globally and locally inhomogeneous: nodes form dense groups inside the networks. These groups are called communities. In this paper we will use the hub percolation community detection method of Bóta et al. to examine and compare the community structure of two word association graphs based on two different languages: English and Hungarian. The English network was created from the University of South Florida word association norms. The Hungarian network was constructed by László Kovács using data collected from Internet users. We examine around which words are communities formed – for example category names or collective nouns. We will also examine if these specific words and their associated groups are similar in the English and Hungarian networks.

Keywords: network science, community, community detection, word associations

MSC: 05C82, 91F20

1. Introduction

The description and analysis of graphs and their properties is a well-established field of science. In recent decades more and more real-life networks were observed, and several common properties of them were discovered. The famous experiment of Milgram [14] demonstrated, that the distance between individuals in a social network is very small. In their paper Barabási and Albert [1] studied the degree distribution of the nodes of the Internet and found, that it follows a power law. The study of infection processes on these networks is also a very popular field with many applications [9]. Finally, the edge distributions is not only globally, but also locally inhomogeneous [12]. This latter property implies that the nodes of these networks tend to form groups or communities, hence this feature of real networks is called *community structure.*

The identification of communities or *community detection* is a popular field of science. One of the main questions of this field is whether a node can belong to more communities at the same time. Traditional or non-overlapping community detection defines communities as disjoint vertex sets based on the following intuition: They are looking for a partitioning of the nodes, which maximizes the number of edges inside the partitions, and minimizes them between the partitions[1]. Community detection algorithms became able to handle large graphs in the last decade. Performance can be further improved by traditional methods or other approaches like graph reduction [2]. An excellent review of community detection can be found in [6].

Overlapping community detection allows nodes to belong to different communities at the same time. The first overlapping community detection algorithm, the Clique Percolation Method (CPM), was introduced by Palla et al. in 2005 [13]. Since then several other methods have been proposed [7, 10]. However, the definition of overlapping communities may change depending on the application. In some papers overlapping community structure requires just a small fraction of communities to belong to multiple groups. [7]. Other applications require a dense, highly overlapping community structure [10].

In this paper, we are going to introduce an application of the overlapping community detection method proposed by Bóta et al. in [4, 3]. We are going to study and compare two word association graphs in different languages. One of them is the well-known network created from the University of South Florida word association norms [11] by Palla et al. [13]. The other one is a Hungarian network constructed by László Kovács using data collected from Internet users. Our main goal is to study the organizing laws behind the formation of communities. Are communities formed around specific words and if so, what characteristics do these word have?

This paper is organized as follows: We will begin with a short introduction of the hub percolation method of Bóta et al. Afterwards we will describe the above mentioned word association graphs. Then we will examine the inner structure of the communities by identifying the nodes that are central to their formation.

2. The hub percolation method

The hub percolation method [3] is a high-resolution clique-based overlapping community detection method created for small-world networks. It has two basic ideas at its core. Consider a graph $G(V, E)$. Fully connected subgraphs of G are called cliques. The identification of cliques is the center of many community detection methods because they represent the ideal community: each member is connected to all other members. The second idea is the observation, that some nodes of a net-

[1] Meanwhile discarding trivial solutions like all nodes belonging to one community.

work are more important than others: they are holding the communities together. These nodes are called hubs.

According to the hub percolation method first we find the set of all maximal cliques C. There are existing algorithms for this purpose; in our works we have used the modified Bron-Kerbosch algorithm introduced in [5]. With the help of C, the set of hubs $H \subseteq V(G)$ is selected. The specific way this is done is governed by the hub-selection strategy of the algorithm. Then we create the set of extended cliques C' from the cliques formed by the hubs using a limited percolation rule. Members of C' can be considered as the building blocks of community detection. Finally, the communities of the graph are created by merging the extended cliques if they contain the same hubs.

The hub selection strategy governs an important part of the algorithm. By changing the strategy it is possible to change some properties of the community structure discovered by the method. In accordance with the idea discussed in the beginning of this section, we assign a value h_v to each node of the network $v \in V(G)$ based on how many maximal cliques does v belong to. We select v as a hub if h_v is greater than some threshold. As general observation [3] we can say, that hubs represent *locally* important nodes of the network. Therefore whether a node is selected as a hub should depend on some t-neighborhood of the node, where t is a small number. In our experience the most effective hub selection strategy is the following: for each node $v \in V(G)$ we count the average hub value in its direct neighborhood ($t = 1$) and multiply it with a parameter $0 < q \leq 2$. We select v as a hub, if h_v is greater then this value.

This parameter allows us to fine-tune the results, and enables the use of the hub percolation method in many different applications. As a general rule we can say, that increasing q decreases the sizes of communities and the overlaps between them. A user should begin with $0.5 \leq q \leq 0.8$, and change its value towards the desired outcome. A detailed description and analysis of this method can be found in [3].

3. Word association graphs

In this section we will introduce two word association networks. The first one contains English words, and is based on the University of South Florida word association norms [11]. Work on this database began in 1978, and it incorporates almost three-quarters of a million associations from 6000 participants. The participants were presented with a discrete association task. They were given a stimulus word, and they had to respond with the first word that came to their mind that was meaningfully related or strongly associated to the presented word. It is easy to see that this structure is a graph. The nodes represent the words, and a directed edge points from one word to another if there is an association from the first word into the other. It is also possible to assign weights to the edges based on the number of associations. The network studied in this paper was created by Palla et al. [13] for the purpose of testing their community detection algorithm. They have created an undirected network by placing undirected edges between the words if there was a directed edge between them in both directions in the original graph. The weight of the new edge was the sum of the two old ones. Then a weight threshold was applied and only edges with weights above this threshold were kept. The resulting network had 7207 nodes and 31786 edges.

The second network was created from the Internet word association database "Agykapocs.hu" by László Kovács [8]. The database collects word associations in Hungarian continuously since 2008. As before, registered users were presented with discrete association tasks. The network studied in this paper was created by omitting the direction of the connections. It has 25431 nodes and 75584 edges.

4. Community structure analysis

Our goal is to identify the laws behind the formation of communities in word association graphs with the help of the hub percolation method described in section 2. More precisely, we are going to examine if communities are centered around specific words. We will call these words as "central" and the identification of them will be our first task.

In order to do this we are going to construct a loose ordering of the nodes of the network with more central words at the top of this ordering. We will denote this ordering as the *global ranking*, and the words of the networks will have a rank property based on their place in this ordering. This is a "loose" ranking, because only the relative position of the words is interesting to us. Small differences in the ranking – e.g. whether they are the 6th or 10th in the word list – are not important for our analysis. After this word list is created some general observations will be made on the global ranking of the vertices of the individual networks.

Finally, we are going to examine if these nodes are indeed central in the communities. After computing the community structure of both networks with the hub percolation method, we are going to construct *local rankings* inside the communities. There are two ways to locally rank the nodes: we can simply order the vertices of the communities according to the global ranking or we can construct an independent ranking on the subgraph induced by the nodes of the community. By comparing the rankings we will show, that the communities found by the hub percolation method are formed around the central nodes of the networks.

4.1. Global ranking

Two communities of the English network can be seen of Figure 1. We can see, that they are centered around category names or collective nouns[2]. Words like these are associated with many other words, therefore it makes sense, that they are the centers of the communities as well. We will adopt this intuition and construct the global ranking according to it.

[2] The first one is centered around the words bird and animal, while the second one is centered around poem and poetry.

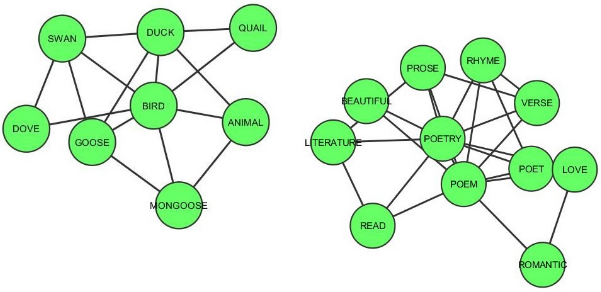

Figure 1: Communities of the English word association graph

In order to continue, we must first decide which words are considered to be “central". Manually deciding this for each word in both graphs would be tedious, but fortunately we can make an observation. According to our hypothesis central words are often associated with other words, therefore it is worthwhile to consider measurements like PageRank or simply the degree of a node and order the vertices of the networks according to them, with the highest values at the top. In our works, we have used the PageRank values, but it is well known, that in undirected networks there is a high correlation between these measurements. If we take a look at the top few hundred words of any of these orderings we can see, that they are exactly the words we are looking for: collective nouns, general adjectives, category names, etc. We can see the 12 highest ranking words from both networks in Table 1.

	1	2	3	4	5	6
HUN	Money	Love	Help	Car	Man	Movie
ENG	Food	Water	Money	Car	Bad	Animal
	7	8	9	10	11	12
HUN	Shock	Economy	Manager	Politics	Nature	Advertisement
ENG	Good	Paper	House	Love	Work	Clothes

Table 1: The highest ranking words in the global ranking

If these words are central to the communities of the networks they must be present in them. For the English network we can say, that 79 % of the communities contain at least one word from the top 300 ranking words. In the Hungarian network the top 100 words are able to cover 95 % of the communities. Despite the differences of the two analyzed networks it is common in them, that the most central words are members of a very large fraction of the communities.

4.2. General observations

We can make several observations on the global rankings of both networks. If we examine the types of the top 200 words we can say, that almost two thirds of them are nouns in both networks. The remaining words are either adjectives or verbs. Fortunately just a small number of the words in the lists is a homonym or a polyseme, so there was no need to apply special rules for these words.

We can draw some interesting conclusions just by comparing the highest ranking words in the English and Hungarian networks. For example, there are only 59 common words in the top 200 highest ranking words in both networks. These common words can be grouped together.

- Basic needs: food, money, car, family, child, etc.,
- Everyday activities: work, travel, sleep, etc.,
- Common adjectives: colors, good, bad, expensive, cheap, etc.

This means that 141 words are different. Some of the differences between the English and the Hungarian lists can be explained by the fact, that the words of the networks were collected at different times (e. g. cell phone, email), on different geographical locations (e. g. Hungary, Europe). Structural characteristics of the two languages (English and Hungarian) can account for the variation in the highest ranking words too. Further sources and causes for the differences will be investigated closely in the future

4.3. Local ranking

Previously we have constructed an ordering of the vertices of the network based on how central they are and found, that the most central vertices are members of almost every community. We did this because we assumed that communities are formed around the frequently associated words. In this section we are going to test this hypothesis. After we have computed the communities of both networks with the hub percolation method, we will examine each community. The global ranking imposes an ordering on the vertices of each community according to their global centrality. We can also construct a *local* ordering by considering the subgraph induced by the vertices of the community, and creating a ranking on the subgraph with the same method.

This way we have two rankings on the nodes of a community: one representing how central they are from a global point of view, another considering only the vertices of the community itself. By comparing these rankings we can decide how well the inner structure of the communities follows our intuition: are the central words of the global ordering also central inside the communities?[3]

[3] That a globally central node has a locally central role too is far from trivial, we can easily consider subgraphs, that do not keep this structure, for example an edge in a star graph.

To compare the rankings we are going to use the rank correlation coefficient of Goodman and Kruskal:

$$\gamma = \frac{n_c - n_d}{n_c + n_d},$$

where n_c denotes the number of pairs of words, that are in the same relative position in both rankings, and n_d denotes the number of pairs, that are in the opposite order[4].

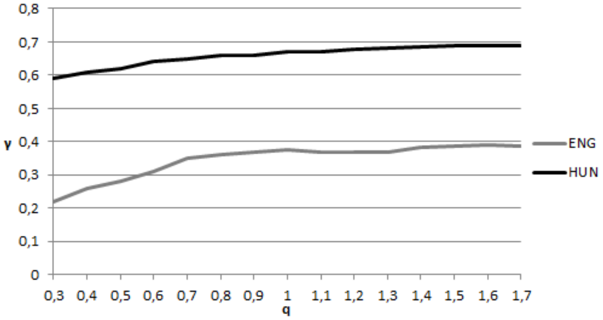

Figure 2: Rank correlation (γ) values for both networks with different values for q.

At this point we are going to use the customizable nature of the hub percolation method and find the setting, that maximizes this correlation. We can see on Figure 2 the averaged rank correlation values for the communities of both networks computed with different values of q, a parameter of the detection method. We can see, that there is a loose positive correlation between the rankings on the English network, while on the Hungarian network this value is much greater. This confirms our hypothesis: although there are differences between the networks, the globally central words are also central to each individual community created by the hub percolation method.

5. Conclusions

In this paper we have used the hub percolation method of Bóta et al. to discover the community structure of two word association graphs in different languages. We have examined the words, that have high degrees and centralities and found, that they are very often category names, common adjectives or collective nouns.

Our main goal was to confirm whether the central nodes of the networks are also central to the communities of them. We did this by constructing a ranking of words both on the whole network and inside communities and then we compared

[4] For example consider two orderings: ABC and ACB. AC and AB are in the proper order, but B and C have different relative positions, therefore the correlation between them is 1/3.

the correlation of these rankings. We have shown, that even though there is a difference between the networks, this correlation exists; it is weaker in the English network and stronger in the Hungarian one.

Acknowledgments. The first author was supported by the European Union and co-funded by the European Social Fund. Project title: “Telemedicine-focused research activities on the field of Matematics, Informatics and Medical sciences”. Project number: TÁMOP-4.2.2.A-11/1/KONV-2012-0073

References

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

[2] Bartha M., Krész M., A depth-first algorithm to reduce graphs in linear time, *Proceedings of the 11th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, IEEE Computer Society, 273–281, 2009.

[3] Bóta, A., Krész, M., A high resolution clique-based overlapping community detection algorithm for small-world networks, Submitted.

[4] Bóta, A., Csizmadia, L., Pluhár, A. Community detection and its use in Real Graphs, *Proceedings of the 2010 Mini-Conference on Applied Theoretical Computer Science* (2010), 95–99.

[5] Eppstein, D., Strash, D., Listing all maximal cliques in large sparse real-world graphs. *Experimental Algorithms*, Springer Berlin Heidelberg, 364–375, 2011.

[6] Fortunato, S., Community detection in graphs, *Physics Report*, **486**(3):75–174, 2010.

[7] Gregory, S., Finding overlapping communities in networks by label propagation. *New J. Phys.*, **12**(10):103018, 2010.

[8] Kovács, L., Conceptual Systems and Lexical Networks in the Mental Lexicon, (In Hungarian: Fogalmi rendszerek és lexikai hálózatok a mentális lexikonban) Tinta Könyvkiadó, Budapest, 2013.

[9] Krész M, Pluhár A., Economic Network Analysis based on Infection Models, to appear in *Encyclopedia of Social Network Analysis and Mining*, Springer, 2014.

[10] Lancichietti, A., Radicchi, F., Ramasco, J. J., Fortunato, S., Finding statistically significant communities in networks. *PLoS One*, **6**(4):e18961, 2011.

[11] Nelson, D. L., McEvoy, C. L., Schreiber, T. A., The university of south Florida word association, rhyme, and word fragment norms, http://w3.usf.edu/FreeAssociation/, 1998.

[12] Newman, M. E. J., Girvan, M., Finding and evaluating community structure in networks, *Phys. Rev. E*, **69**(2):026113, 2004.

[13] Palla, G., Derényi, I., Farkas, I., Vicsek, T., Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**(7043):814–818, 2005.

[14] Travers, J., Milgram, S., An Experimental Study of the Small World Problem, *Sociometry*, **32**(4), 425-443, 1969.