ACTA PHYSICA DEBRECINA XLVII, 57 (2013)

EXAMINATION OF AN EDGE WEIGHTED NETWORK CREATED FROM DATA OF COUCHSURFING

E. Fenyvesi

University of Debrecen, Department of Experimental Physics, 4032 Debrecen, Egyetem Tér 1., Hungary

Abstract

Couchsurfing is an example of evolving self-organized social network. Members of the Couchsurfing network offer hospitality exchange and social networking services for other members. Significant amount of data is available that can be used for analyzing the system with network science tools. This paper presents some results of an analyzation of a network created from a randomly selected part of data from Couchsurfing. The nodes in the network represents countries visited by the users of Couchsurfing, and two countries are connected if at least one of the users visited both two countries. Weight of an edge shows the sum of the users who visited the countries connected by the given edge. Edge weight distribution of the network was investigated, and it revealed that the distribution decays as a power law. The network's community structure was explored by using fast greedy modularity optimization algorithm [1], and a dendrogram was created to visualize the results.

I. Introduction

Various systems (which imply a large number of interacting elements) can be analyzed by using the improving methods of network science. Theories and methods of mathematics, physics and computer science are used to

study these complex networks, which can be modeled with graphs, where nodes represents the elements of the network, and edges show the interactions between them. Various types of quantites can be assigned either to nodes and edges (node- and edge weight, edge direction, etc.), depending on the properties of the system to be examined. Considering from these quantities as much as possible helps us to get more knowledge of the network. For example, in the case of world-wide airport network, nodes represent airports and edges represent direct flights between them. Edge directions and weights show the number of flights from one airport to another during a given period of time. (If only an unweighted and undirected edge connects two airports, it only means that there exist a flight between them.) Algorithms were invented to characterize the role and importance of nodes (for example their centrality), and to find the communities formed from them. In this paper a network created from data of Couchsurfing users is examined by using `gephi`[3], `igraph` [1] and other programs developed by the author.

II. Couchsurfing

According to Wikipedia "Couchsurfing International Inc. is a Delaware C corporation based in San Francisco that offers its users hospitality exchange and social networking services"[2]. Registration on couchsurfing.org is free. After registration, users have their personal site, where they can take connection with people from different countries, find a host for free when they want to travel to a given country, accommodate people from other countries, and load information of themselves including the names of countries they traveled to. A network can be created from this data, where nodes are countries and an undirected edge is present between any pair of countries if at least one of the users has traveled to both countries. Weights can be assigned to edges, the weight of an edge connecting a given pair of countries is equal the the sum of the users who have traveled to these two countries. This network can be analyzed with the tools of network science.

III. The network

In this paper data from 600 users was used to create the network of coun-

tries. The original network included $N = 150$ nodes and $E = 5997$ edges. This network was highly interconnected and very dense ($D = 0,537$). Density is equal to the actual number of links in the network divided by the number of all possible links: $D = \frac{2E}{N(N-1)}$. In order to get more relevant information of the system, density needed to be decreased by one magnitude. After deleting edges which have weight $w < 6$ from the network, density decreased to $D = 0,045$. The new network includes $N = 66$ nodes (countries) and $E = 501$ edges (Fig. 1). Degree of a node is equal to the number of edges connected to it. In our case, average degree shows that $6,68$ visitors traveled to a given country in average. Average path length (a) is calculated by finding the shortest path between all pairs of nodes, adding them up, and then dividing by the total number of pairs. This shows us, on average, the number of steps it takes to get from one node to another. In our case $a = 1.9613$. Diameter (d) is the longest of all the calculated shortest paths in a network. For our network, $d = 5$.

IV. Edge weight distribution

Edge weight distribution decays as a power law: $P(w) \propto w^{-\gamma}$ with $\gamma = 2.833$ (Fig. 2.). The distribution is scale invariant, what means it does not change if the scale of weight is multiplied by a common factor: $P(cw) = b(cw)^{-\gamma} = c^{-\gamma}P(w) \propto P(w)$. Pareto law [5] holds for quantities with power law distribution. For our network, it means that 80% of total weight of edges corresponds to 20% of the edges with largest weights $w > 12$. After deleting edges with $w < 13$ the remaining network includes 24 nodes. The remaining countries are 36.36% of the initial 66.

V. Closeness centrality of edges

The closeness centrality of a vertex measures how easily other vertices can be reached from it (or the other way: how easily it can be reached from the other vertices). It is defined as the number of vertices minus one divided by the sum of the lengths of all shortest paths from/to the given vertex. Closeness centrality reflects the popularity of countries among Couchsurfing users. On Fig. 1. size of the nodes are proportional to their closeness

centrality. The most central countries are: France, England, Germany, Italy, Spain. Force-directed graph drawing was applied to draw nodes with greater centrality to more central positions in the drawing. It can be noticed, that the most central countries are Western European countries. Australia, Asian and South American countries are less central.

VI. Communities

Community structure is the gathering of vertices into groups such that there is a higher density of edges within groups than between them. [4] Pairs of countries are more likely to be visited by the same person if they are both members of the same community, and less likely to be visited if they do not share communities. `Igraph` includes a function for finding community structure, which uses fast greedy modularity optimization algorithm [1]. Another function draws a dendrogram, showing community structure of the network (Fig. 3.) We expected countries on the same continent to be in the same community, because if somebody wants to go often on short and inexpensive holidays, in the most possible case he/she chooses countries on the same continent where he/she lives. But if somebody wants to go on a long holiday, and has enough money to travel to another continent, it is possible that he/she visits more than one country on the same continent. The expectation is right in general, but there are interesting exceptions. It can be noticed that Central European countries (Czech Republic, Slovenia, Slovakia, Hungary, Latvia, Croatia, Poland, Austria) form a community. But Switzerland, which is also a Central European country, is in a small community with Greece, Costa Rica, United States, Canada, and Mexico. It is not surprising that the last four ones are in the same community, since they are North American countries. Far Eastern countries, like Laos, Vietnam, Cambodia, Thailand, China, Malaysia, and Indonesia form another community, but surprisingly Singapore is in the community that collects Central European countries. Bolivia, Uruguay, Argentina, Brazil, and Chile are all in the same small community of Southern American countries. Nepal, Tunisia, Serbia, Cuba, France, Finland, Israel, and Jordan forms an interesting mixed community with countries from four different continents. Also interesting to see the structural properties of the dendrogram. It seems like the program finds a pair of countries as the "smallest" community and then

links a third country at one level above the pair, and a fourth at an upper level and so on. Communities formed in this way are linked together only at the highest levels of the dendrogram. But there are a few interesting exception, where pairs or small communities are linked together at low levels. That happens in the case of France, Finland, Israel, Jordan, or England and Scotland with the Far Eastern countries. On Fig. 1. countries are collected into the four main clusters, and the membership is marked with different colours. It is interesting to see that countries with the highest centralities are in different communities, and every community contains central and also peripherical countries (except for the yellow community with South American countries).

VII. Discussion

In this paper it was shown that an edge weighted network can be created from the data of Couchsurfing social newtwork. The investigation of this network explored that edge weigh distribution decays as a power law with exponent $\gamma = 2.833$. Organizing processes resulting this distribution is a subject of further investigation. Centralities of nodes were also examined, and it turned out that the most central countries are France, England, Germany, Italy, Spain. Community structure of the system reflects the traveling habits of Couchsurfing users. Countries on the same continent tend to form a community, it means if somebody travels to more than one country, in the most possible case he/she chooses countries on the same continent.

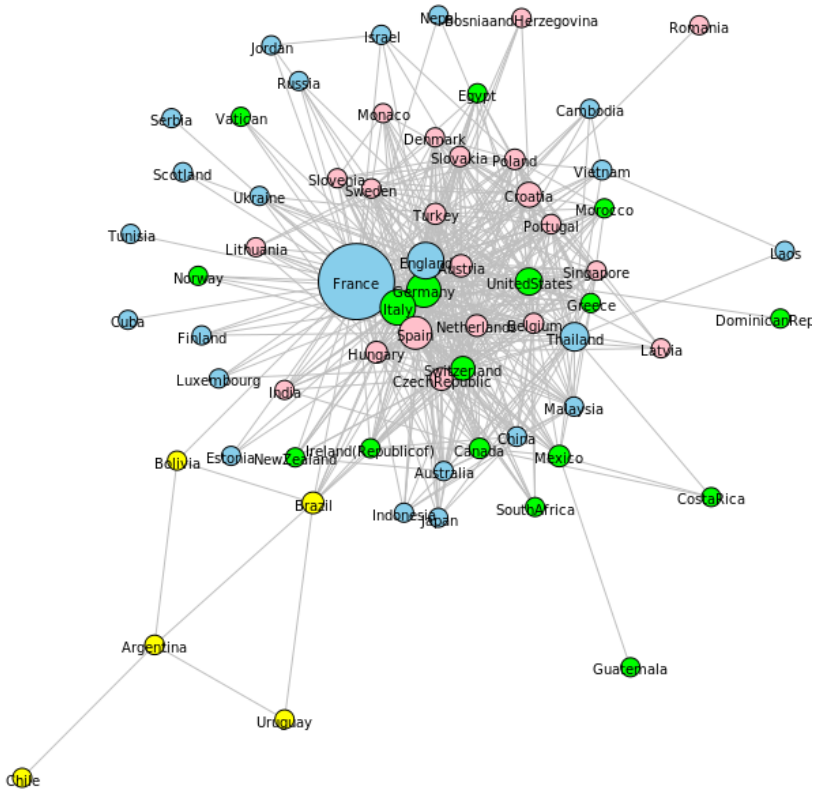

Figure 1: The network. Colours of nodes indicate their community membership, while size of them shows their centrality.

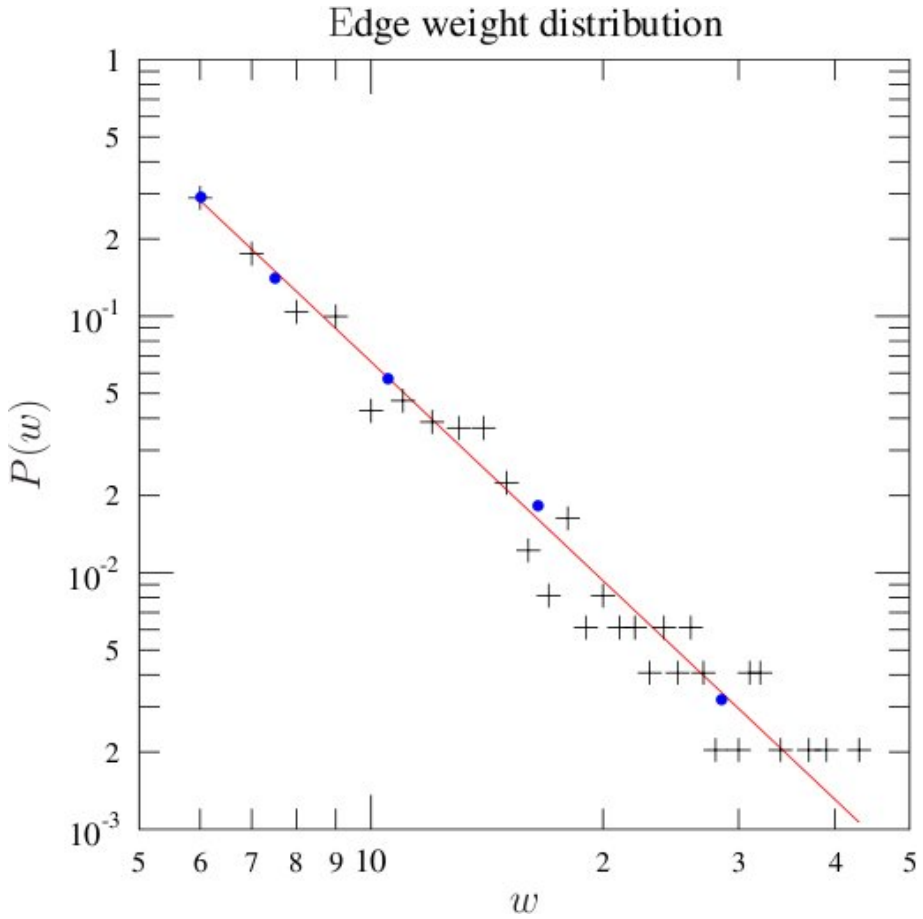

Figure 2: Edge weight distribution. "Logarithmic binning" was applied because the data becomes noisy at higher values of edge weights. This happenns because the number of samples in the bins becomes small and statistical fluctuations are therefore as large as a fraction of sample number. The straight line was fitted to "logarithmic binned" data.

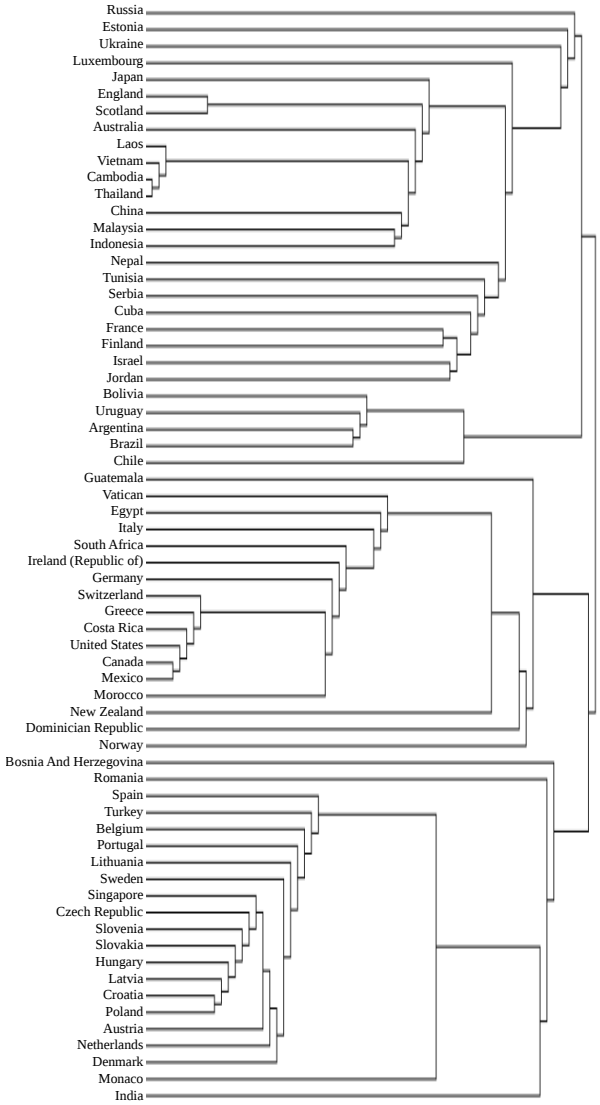

Figure 3: This dendrogram illustrates the arrangement of the communities produced by fast greedy modularity optimization algorithm [1].

References

[1] Gabor Csardi and Tamas Nepusz InterJournal, **Complex Systems**, 1695 (2006)

[2] http://en.wikipedia.org/wiki/CouchSurfing

[3] http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[4] A Clauset, MEJ Newman, C Moore Phys. Rev. E **70**, 066111 (2004).

[5] Pareto, Vilfredo; Page, Alfred N. *Translation of Manuale di economia politica ("Manual of political economy")* (A.M. Kelley, 1971).