A Network Science Perspective of a Distributed Reputation Mechanism

Rahim Delaviz, Niels Zeilemaker, Johan A. Pouwelse, and Dick H.J. Epema

Delft University of Technology, the Netherlands
Email: r.delavizaghbolagh@tudelft.nl

***Abstract*—Reputation mechanisms are widely used in online networks to rank users or products, but despite their importance, very few studies have been done or published on their real behavior. In this paper, we study an Internet-deployed distributed reputation mechanism called BarterCast that is specifically designed for peer-to-peer file-sharing systems. The BarterCast mechanism is based on building a weighted directed graph from the data transfers that have occurred among the peers, and on employing the Maxflow algorithm in this graph to evaluate reputations. In this paper, we study this mechanism from the network perspective and we provide a detailed analysis, which includes such network topology measures as the degree distribution, node interconnectivity, the clustering coefficient, community structure, and distance measures. Besides, we study the geographical spread and content sharing behavior of the system participants and correlate the results with their connectivity in the network. We interpret each evaluated measure in the scope of reputation and file-sharing mechanisms and propose relevant implications and prospective applications for future designs. All the measurements are based on data that we have collected during two years of crawling the Tribler file-sharing network, which employs BarterCast as its reputation mechanism.**

I. Introduction

Despite much interest in reputation systems over the last few years, there are hardly any studies of the real behavior of Internet-deployed decentralized reputation systems. In this paper, we study the Internet-deployed distributed reputation mechanism called *BarterCast* [1], which builds a graph structure for its operation, from a *network science* perspective, and employ many network measures to understand its behavior. BarterCast is used in the BitTorrent-based peer-to-peer content-sharing system Tribler [2] to rank the peers based on their sharing behavior. In this study, using a number of datasets from a real operational environment, we build a network of content-sharing activities. Employing this network, we calculate a number of appropriate measures to comprehend the network structure and the operational behavior of the underlying reputation mechanism. We interpret each calculated measure in the scope of reputation mechanism or content-sharing system and provide an explanation of its implication. For some of the measures we elaborate on their prospective applications for further improving the reputation mechanism.

In BarterCast, peers exchange messages about their upload and download actions, and they use the collected information to evaluate the reputations of other peers. From the BarterCast messages it receives, each peer builds a local weighted directed graph with nodes representing peers and edge weights representing amounts of transferred data. This subjective graph is then used by each peer to calculate the reputation values of other peers by applying the Maxflow algorithm to the graph, interpreting the edge weights as “flows”.

To collect the required data, we started a Tribler network crawler in September 2010, which still is running; Tribler is a BitTorrrent-based peer-to-peer file sharing client that is used for peer-to-peer file-sharing and video-on-demand services [2], and that uses BarterCast to rank peers. The main task of the crawler is to discover peers and to collect data transfer records from them. Using the permanent identifiers of the peers we are able to correctly group the collected records from different peers and to generate a global network, which we call the *work-graph*. Moreover, in the Tribler network, there are four so-called *super peer* nodes which are used for bootstrapping. We employ the data recorded by these super-peer nodes to generate two sets of valuable information about peers. The first of these sets contains the IP addresses of the peers, which enable us to do a geospatial analysis of the network. The second contains the content swarms peers have participated in, which allow us to perform a content-based similarity analysis of neighbor peers in the network.

We perform an analysis of the topological characteristics of the work-graph, which include the *degree distribution*, the *nodes interconnectivity*, the *clustering coefficient*, the *community structure*, and *centrality* and *distance* measures. The degree analysis of the work-graph shows that, like social network graphs, it has a long-tail power law distribution. We test the hypothesis of it being obeying a power law and check it against similar distributions. The interconnectivity analysis shows that the graph has star-shape structures and is far from a scale-free graph; this result is confirmed by the low clustering coefficient of the graph. Complementary to clustering coefficient, it is observed that the graph has strong communities. Moreover, we observe that there is a strong correlation between the node degree and the betweenness and closeness centrality measures. This observation suggests that node degree is a good approximation for these complex measures. Finally, using the temporal graphs that we build over time, we can observe how the diameter, the average path length, and the density of the graph change over time. These measures are covered in Section IV.

Complementary to the general topological analysis, in Section V, we present results on the geographical spread of nodes at the granularity of ISPs, and we evaluate the correlation of ISP co-location and having an edge (content sharing) in the graph. Finally, in Section VI, we present our findings about user upload and download behavior and measure the content-based similarity of neighbor nodes in the graph. These similarities are based on the complementary data that we have about the content that peers have shared with others.

II. Related Work

For years, social scientists have studied human relations and have made interesting discoveries, e.g., the small-world phenomenon by Milgram [3] and the partitioning social relations into strong and weak ties by Granovetter [4]. Recently, due to the fast growth of online social networks, researchers have put a lot of effort in studying the static and dynamic properties of these networks. Kumar et al. [5] have studied friendship relations in Flickr and Yahoo360, and they have shown that these networks have a large Strongly Connected Component (SCC). An analysis and comparison of the social networks of Flicker, YouTube, LiveJournal, and Orkut by Mislove et al. [6] confirms the power law, small-world, and scale-free properties of these networks. Recently, the Facebook network, due to its high popularity and size, has attracted many researchers. Orthogonal to other studies, which are more focused on general network properties, Viswanath et al. [7] and Wilson et al. [8] have studied this network from the user-activity and link-reliability perspectives. They have clarified that despite the high number of links (friendship relations), only a small portion of the links of a node are reliable, meaningful, and useful for real-life applications.

Besides social networks, analyzing the structure of the Web and Internet links has led to many interesting findings as well. Analysis of the *Autonomous System* (AS) level of the Internet by Mahadevan et al. [9] has revealed that the *joint degree distribution* can characterize Internet connections. Falatous et al.[10] have shown that the Internet topology follows a power law degree distribution; a claim that has raised criticism as well [11]. Besides the study of general network structures, there are some studies on the geographical properties of the Internet infrastructure [12], [13], which study the geographical spread and node distances. Regarding the Web network, a study of the Web links by Broder et al. [14] shows that links have a “bow-tie” shape, with a large SCC and many small groups of nodes connected in one-way to SCC.

Researchers have proposed wide applications of social networks in other systems, e.g., defense mechanisms [15], recommendation systems [16], reputation systems [17], and many others. Despite many proposals only a few of them have gone beyond design into a real application, and even for those who reached that level, there is no real large-scale study of their behavior. Our study is distinguishable from similar works in two folds. First, we perform a thorough analysis of a large scale and deployed mechanism from a network perspective. Second, despite pure social network studies, which only present a number of general measures, we look at the calculated measures from the reputation mechanism perspective and provide valuable hints for future designs.

III. The BarterCast Mechanism

The BarterCast mechanism belongs to a class of peer-to-peer incentive mechanisms where a contributing peer is rewarded by other peers in the network and direct compensation is not expected. This mechanism is used in the Tribler BitTorent client to rank peers according to their upload and download behavior. In this mechanism, a peer whose upload is much higher than its download gets a high reputation, and other peers give a higher priority to it when selecting a content bartering partner. In BarterCast, when two peers exchange content, they both log the cumulative amount of transferred data since the first data exchange along with their identities in a BarterCast record. In Tribler, peers regularly contact other peers in order to exchange BarterCast records.

From the BarterCast records it receives, each peer creates its own current local view of the upload and download activity in the system by gradually building its *partial graph.* The partial graph of peer i is the weighted directed graph $G_i = (V_i, E_i)$, where V_i is the set of peers whose activities peer i has been informed about through BarterCast records, and E_i is the set of edges (u, v, w), with u, $v \in V_i$ and with w the weight representing the total amount of data transferred from u to v. Upon the receipt of a BarterCast record (u, v, w), peer i adds the edge $u \to v$ to G_i if it did not exists, otherwise it updates the weight of this edge.

To calculate the reputation of an arbitrary peer $j \in V_i$ at some time, peer i applies the maxflow algorithm [19] to its current partial graph to find the maximal flow from itself to j and vice versa. Maxflow is a classic algorithm in graph theory for finding the maximal flow from a source to a destination node. In the original Maxflow algorithm, all paths are considered in carrying flow, but in BarterCast paths longer than h are ignored. This limited version of the algorithm is called *h-hops* Maxflow. When applying Maxflow to the partial graph, we interpret the weights of the edges as flows. If $\Phi_h(x, y)$ is the h-hops maxflow from x to y, then the *subjective reputation* of peer j at peer i is calculated as:

$$R_i(j) = \frac{\arctan(\Phi_h(j,i))}{\pi/2} \times (1 - \frac{\arctan(\Phi_h(i,j))}{\pi/2}), \quad (1)$$

and so $R_i(j) \in [0, 1)$. If the destination node j is more than h hops away from i, then its reputation at i is zero.

In our previous work [18], we presented a dissemination mechanism for BarterCast records, that provides nodes a near-complete view of the generated records in the network. Applying the targeted dissemination the partial graphs converge to a graph that contains all the edges and nodes. In this paper, we base our analysis on this graph that is called the *work-graph*.

IV. Topological Characteristics

In this section, we study the work-graph of BarterCast from the network topology perspective, and we present a number of

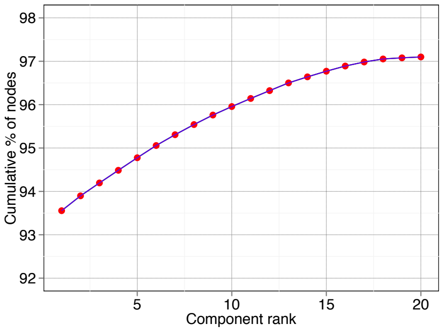


Fig. 1. The cumulative coverage of the connected components ranked according to decreasing size.

relevant measures that help us to understand the connectivity pattern of the nodes. For ease of reading, the terms graph and network are used interchangeably.

A. The Undirected Work-graph

In the BarterCast work-graph, an edge indicates the amount of data transferred from one peer to another, but from the interaction perspective, its direction is not important. So, for the following analysis we remove the edge directions, and we add the weights if there are edges in both directions between two nodes. The original directed graph contains of 73,201 nodes and 352,042 edges; after removing directions, the number of edges is 283,973. In Section IV-C, we present some measures on the edge and weight symmetry, but unless stated otherwise, our analysis is based on the undirected work-graph.

Furthermore, since most of the graph measures, like the clustering coefficient, only make sense when the underlying graph is connected, we consider the *Largest Connected Component* (LCC) of the work-graph. In total there are 939 connected components, out of which 780 contain only two nodes. Figure 1 plots the cumulative percentage of the nodes covered by the largest 20 connected components ranked according to decreasing size. As the plot shows, 93.55% of the nodes belong to the LCC. The number of nodes and edges in the LCC are 68,315 and 265,033, respectively. In conclusion, since the LCC is a good representative of the whole graph, we base our analysis on the LCC unless stated otherwise.

B. Degree Distribution

The original work-graph is directed, and in such a graph a node has three types of degrees: the *in-degree*, the *out-degree*, and the *total degree*, which is the sum of in-degree and out-degree. Figure 2 shows the degree-frequency plot for these three types of degrees. Visually, after a threshold degree of about 30, the plot looks like a straight line. Based on this observation some researchers conclude that such a distribution follows a *power law*, and interpret the graph as a *preferential attachment* graph. But for two reasons this method is not a reliable way to conclude that a distribution follows a power law. First, due to high data (node degree) diversity, many values appear once and the frequency values are not informative. Secondly, using the frequency plot, non-power law driven data, e.g. exponential, can be misleadingly interpreted as power law [20]. Due to these limitations, using the CDF or the Complementary Cumulative Distribution Function (CCDF) is more common [20], [21], [9], [6].

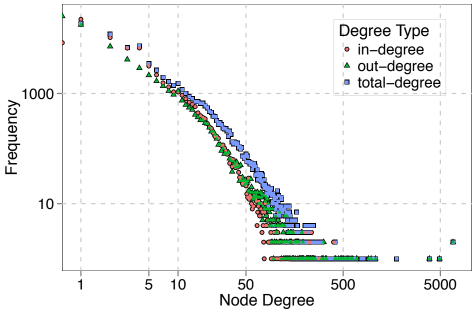


Fig. 2. The degree frequency of the LCC of the work-graph (both axes are in log-scale).

Figure 3 shows the CCDF of the node degree in the LCC of the work-graph. In this plot, it looks as if the tail of the plot follows a power law distribution. A distribution is power law if it is driven from $p(x) \propto x^{-\alpha}$, where α is a fixed value called the *scaling* parameter. Using the method proposed by [22], we estimate the parameters α and x_{min}, where only for values larger than x_{min} the power law holds. To estimate the parameter α, first x_{min} is fixed at some value, and then using *maximum likelihood estimation* and assuming that the data are driven by a power law distribution, the value of α is estimated. To find x_{min}, starting from the lowest possible value, the Kolmogorov-Smirnov distance between the data and the estimated distribution (for the selected x_{min}) is calculated. The x_{min} that gives the lowest distance is chosen as the best value. Applying these methods we estimate $\alpha = 2.88$ and $x_{min} = 42$.

So far, we were able to fit a power law distribution to the degree values and to estimate its parameters, but whether it is a good fit or not is still a question. To evaluate the quality of the fit, using the estimated values for α and x_{min}, we calculate the *p-value* of the *goodness-of-fit* test for power law, and compare

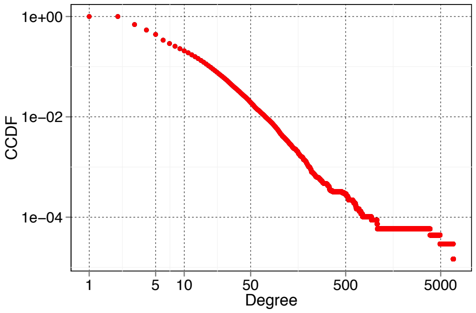


Fig. 3. The complementary cumulative distribution function (CCDF) of the total degree in the LCC of the work-graph (both axes are in log-scale).

TABLE I
VUONG LOG LIKELIHOOD RATIO TEST RESULTS

	Poisson	Log-normal	Exp.	Powerlaw+cut off	Yule
$\mathcal{R}$	+2.78	+0.008	+3.30	+9.03e-6	-0.97
p	**0.005**	0.993	**0.001**	1	0.330

it with a threshold value. For the degree values the obtained p-value is 0.067. In the conservative approach [22], the p-value threshold for rejecting the power law hypothesis is set to 0.1, but generally in a more lenient approach it is set to 0.05. In conclusion, in the conservative approach the hypothesis that the work-graph has a power law degree distribution is rejected, but with the lenient approach this hypothesis is not rejected.

We now further analyze whether with the lenient appraoch, other types of distributions, e.g., the exponential distribution, give a better fit than the power law or not. The *likelihood ratio test* is a simple test for comparing the likelihoods of a dataset of belonging to a number of distributions. The sign of the logarithm of the ratio of two likelihoods, $\mathcal{R}$, can determine which distribution is a better representative for the given data. In practice, relying just on the sign of $\mathcal{R}$ is subject to random fluctuations around zero. To make a solid decision we use the method of Voung [23] that gives a p-value on the significant of the sign of $\mathcal{R}$, for small p-values the hypothesis that the sign of $\mathcal{R}$ is due to random fluctuations is rejected, and vice-versa. Table I presents the results of comparing the power law with four other distributions; a positive $\mathcal{R}$ indicates that the power law should be favored over the other distribution. As the table shows, the power law is reliably favored over the Poisson and Exponential distributions. For Yule, it seems that it is better than power law, but like the Powerlaw+cut off the sign of $\mathcal{R}$ is not reliable.

Summary & Implication: Our analysis of the degree distribution of the work-graph shows that it has a long-tail distribution. Depending on the application behind the network, having a high degree can have different reasons. For example, in a Web network, the popularity of a site can be the main reason for having many links to it. In our work-graph, having only a few very high-degree nodes means that a few peers are responsible for most of the content sharing in the network. Indeed, these are peers who stay online for a long time and are discovered by other peers more often. On the other hand, many low-degree nodes indicate the presence of many short-time users or even *free-riders*. A study of why there are so many short-time users will help to increase the quality of service in the whole network. Finally, a non-random structure means that the network is vulnerable to targeted strategic attacks on highly connected nodes. If an attacker provides the highly connected nodes with a contaminated content, then the content is spread very fast in the network. This is a concern that should be taken into account in future designs.

C. Node Interconnectivity

The degree distribution provides information on the individual connectivity of the nodes but it does not provide information on the relation between the degrees of neighboring nodes. In this section we provide some results on one-hop connectivity of the nodes as captured by the *Average Neighborhood Degree*, the *Assortativity*, and the *Rich Club Community* (RCC) metrics.

Consider the $k \times k$ Joint-Degree Distribiution (JDD) matrix $M = (m_{ij})$, where k is the largest node degree and m_{ij} is the number of edges that connect nodes with degree i to nodes with degree j. Dividing M by the total number of edges gives the probabilities that a randomly selected edge connects nodes with degrees i and j. For large and sparse graphs, the JDD matrix is highly sparse and not very informative. Instead, the average neighbor degree of the nodes of degree x, $k_{nn}k(x)$, is a more informative statistic for sparse graphs. An increasing $k_{nn}k(x)$ is an indication of the tendency of higher degree nodes to connect to other higher degree nodes and vice versa. We plot $k_{nn}k(x)$ in Figure 4, where due to its decreasing trend it seems that higher degree nodes tend to connect to lower degree nodes. A similar but more summarized metric than $k_{nn}k(x)$ is the *degree assortativity* of the graph, which takes values between -1 and +1, values close to +1 indicating the tendency of similar degree nodes to connect to each other and vice versa. For our graph, the degree assortativity is -0.062.

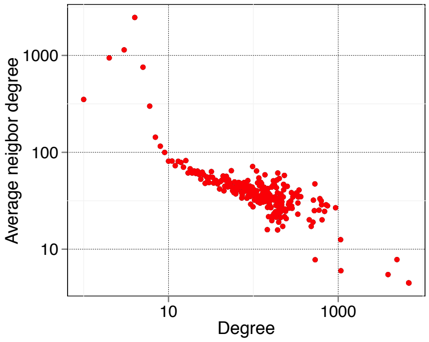


Fig. 4. The average neighbor degree of the graph (both axes are in log-scale).

The last metric for evaluating the connectivity pattern of the nodes is *densely connected core* or *rich club community* [9]. A core is defined as a small group of well connected nodes that connect the remaining nodes. In order to understand the importance of the core nodes, we do a similar experiment as Mislove et al. [6]. In this experiment, we remove a number of the highest-degree nodes (a rich club) from the LCC and count the resulting number of disconnected components; the higher this number, the higher the importance of the removed nodes. Figure 5 presents the fragmentation results of removing different fractions of the high-degree nodes; it shows the cumulative percentage of nodes included in the components ranked according to decreasing size, with components of the same size having the same rank (and counted multiple times in the coverage). Especially for small sizes, there may be multiple components. In this figure the right-most point of each curve represents the single-node components. As can be observed, for every removal ratio, almost all nodes either are part of the LCC or they become single-node components. Such a phenomenon occurs when there is a high number of star-shape structures with the removal of the central, high-degree node leaving many single-node components.

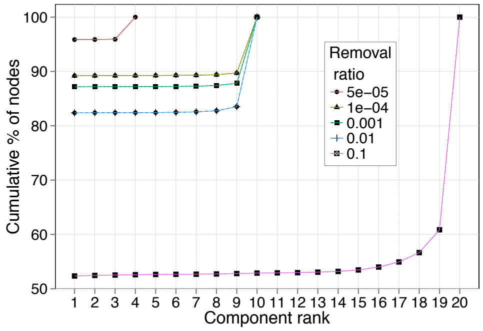


Fig. 5. The rich club community removal effect.

Summary & Implication: The $k_{nn}k$ measure has a decreasing trend, which means that lower degree nodes and new comers tend to connect to nodes that have many links. This trend is similar to user connectivity in YouTube [6], which according to the authors is due to the "celebrity" effect, where popular users have many followers. A similar interpretation holds for the Tribler network as well, since users with many links have more content to share with others and they are more often discovered by those who look for content. Also, a decreasing $k_{nn}k$ means a low likelihood of having a *scale-free* graph [20]. This finding is confirmed by the degree assortativity which similarly to the degree assortativity of the Internet and Web networks is negative [24]. Notice that every scale-free graph is power law but not vice versa.

Finally, the RCC analysis shows a connectivity of the network that is very resilient against the removal of high-degree nodes. For example, in the extreme case of removing 10% of the highest-degree nodes, still more than half of the nodes remain in the LCC, which is in contrast to hub-like graphs where highly connected nodes play a critical role in connecting nodes. In conclusion, it seems that there are some strong community structures in the network, and many nodes are gathered around a few nodes.

D. Clustering & Communities

The degree distribution of a graph indicates the local connectivity of nodes, and the average neighbor degree $k_{nn}k$ gives information on the connectivity of similar degree nodes, but neither gives information on how the neighbors of a node are connected among themselves. In this section we provide results on the *local clustering coefficients* and the *global clustering coefficient*, which indicate whether the neighbors of nodes are tightly connected or not. Figure 6 shows the average clustering coefficients of nodes with the same degrees in the whole graph. The global clustering coefficient of the graph, which is the average of all clustering coefficients, is 0.0066.

Besides the clustering coefficient, we can look for communities in the graph. A *community* is simply a group of nodes with high internal and low external connectivities [25]. Depending on the application behind the network, there can be different reasons for the formation of communities, for example, geographical locality, similar taste, and etc. Since the number and the structure of communities are not known in advance, we need a way to evaluate the quality of a group of communities. Newman et al. [26] have introduced the concept of *modularity*, which quantifies the quality of partitioning a graph into communities. This measure is defined as:

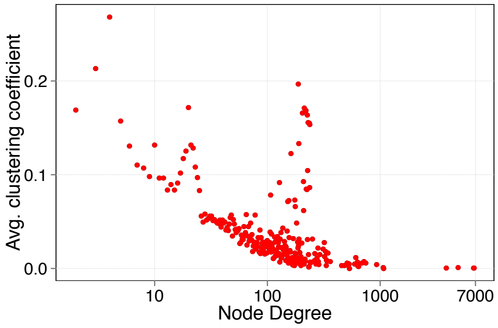


Fig. 6. The average clustering coefficient vs. the node degree (horizontal axis is in log-scale).

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j), \tag{2}$$

where m is the number of edges, A is the adjacency matrix, k_i is the degree of node i, c_i is the community id of i, and δ is the Kronecker delta function. For modularity values close to zero, the partitioning is meaningless, and values between 0.3 and 0.7 are reasonable for quality partitioning [27].

Finding community structures is closely related to the notion of graph partitioning and hierarchal clustering, and there are numerous algorithms for detecting communities [28], [29]. The proposed algorithms mainly vary in their computational complexities, and only a few of them are appropriate for graphs of the size of our work-graph. In this paper we use four algorithms for detecting communities:

- *Fast Gready* (FG) [27], which is an efficient implementation of the hierarchal edge-betweenness community algorithm of Newman et al. [25].
- *Multi Level* (ML), which is based on local optimization of the modularity measure around a node [30].
- *Spin Glass* (SG), which is a simulated annealing heuristic method to optimize the modularity measure [31].
- *Label Propagation* (LP), which is a near-linear algorithm. First, it uniquely labels the nodes, then updates them by majority voting among the neighbors of a node [32].

Figure 7, presents the induced community graphs along with the number of communities and modularity values, obtained by applying the above algorithms. In these graphs, each node represents a community and an edge exists between two nodes if there is at least one inter-community edge, i.e., an edge between nodes in the work-graph from either community. The size of the nodes and the width of the edges correlate with number of nodes in each community and the number of inter-community edges, respectively. As can be observed, the FG

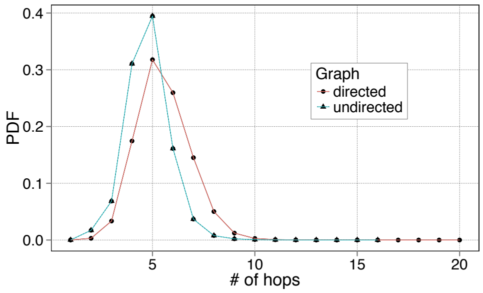


Fig. 8. The density of the length of the shortest paths.

algorithm gives the highest modularity measure but the number of communities is too high. Considering both modularity and the number of communities, the ML method gives the best partitioning.

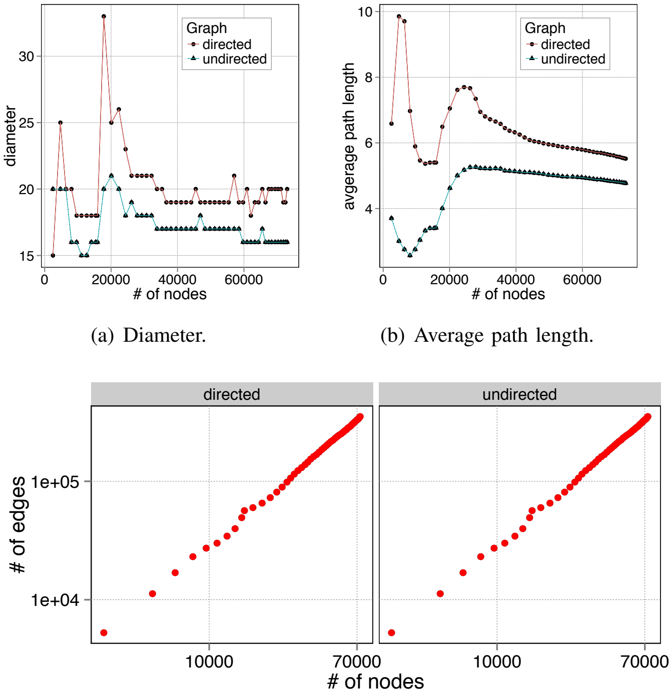
Summary & Implication: The decreasing trend of the clustering coefficient in Figure 6 confirms the previous finding of decreasing $k_{nn}k$, Section IV-C, that high-degree nodes play a crucial role in connecting many of their neighbors. Due to this phenomenon, the data exchange and presence of the high degree nodes is important for the operation of the system.

Regarding the community structures, the main reason for the formation of communities is that the work-graph grows in time, and the nodes that belong to a community are those nodes that were active in a specific period of time. When the time passes, most of the peers leave the network, but a few of them continue on sharing content with the new peers. In network terminology, these peers act as a bridge between one community to the next one(s), and since the number of such long term active peers is not high they are not strong enough to merge communities.

E. Distance Properties

We will now investigate the distance characteristics of *average path length* and *diameter* of the work-graph. Figure 8 presents the probability density of the lengths of shortest paths. It has a mean of 4.83 and 5.52 for the undirected and directed graphs, respectively. In order to understand how these measures change over time, we build temporal graphs based on edges appearing in the network and calculate these measures for them. A temporal study help us to understand how these measures change over time and whether the concept of *densification* [33] holds or not.

To measure the temporal features, starting from the first day of crawling we build temporal graphs in different periods, where the nodes and edges discovered in period n are added to the graph of period $n-1$. In our experiment, for the total crawling period of two years we divide the records in 52 sets of biweekly periods, and for each period we build a graph. Figure 9 presents the diameter, the average path length, and the density of the directed and undirected versions of the temporal work-graphs, which are plotted against the number of nodes. The density of a graph is the ratio of the number of edges to the number of possible edges. As can be observed, despite minor fluctuations, all these measures show a decreasing trend over the long term. Besides, the undirected version of each temporal graph has a lower diameter and average path length than the directed version, but such a difference is not visible for the density. The reason for this phenomenon is that there are relatively few double edges between a pair of nodes than the total number of edges, and when the direction is removed, the density is not much affected.



(a) Diameter.

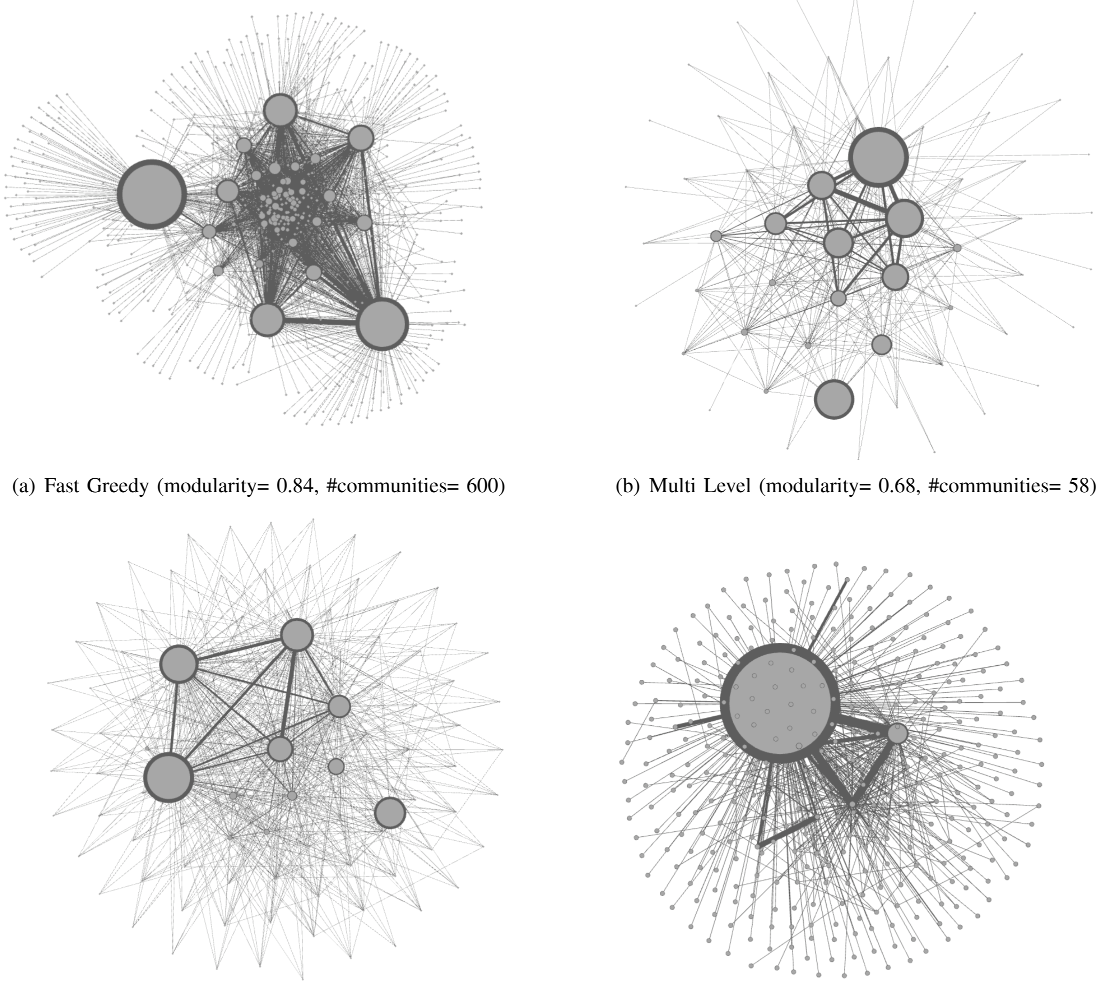
(b) Average path length.

(c) Log-log plot of the number of nodes vs. number of edges.

Fig. 9. The diameter, the average path length, and the density values of the temporal graphs vs. the number of nodes (In Figure 9(c), since the difference between the number of edges is not high we observer similar shapes).

Summary & Implication: The average shortest path of the work-graph is very close to those of online social networks [6], [21] and the AS-level Internet network [9]. Knowing the average path length is useful, since in the reputation evaluation process of BarterCast, the number of hops number in the Maxflow algorithm can be based on it. In this case, since in the final graph the directed average path length is 5.52, in the Maxflow algorithm the number of hops can be set to 5 or 6.

LIke the graphs studies by Lescovec et al. [33], where they have discovered *densification effect* on the studied graphs, we also observe a similar behavior in our graphs. Figure 9(c), plots the number of nodes versus the number of edges in the temporal graphs in log-log scale. As it is observed there is linear correlation between the log of the these two values, and a linear regression fit shows the slope of 1.26, which is greater than 1, and it means that the average degree of the graph increases over time. Moreover, like the studied graphs by Lescovec et al. [33], we observe that the diameter and the average path length have a decreasing trend.



(a) Fast Greedy (modularity= 0.84, #communities= 600)

(b) Multi Level (modularity= 0.68, #communities= 58)

(c) Spin Glass (modularity= 0.59, #communities= 100)

(d) Label Propagation (modularity= 0.23, #communities= 335)

Fig. 7. The community-induced graphs obtained through applying different community detection algorithms (nodes represent communities and edges indicate inter-community edges in the work-graph).

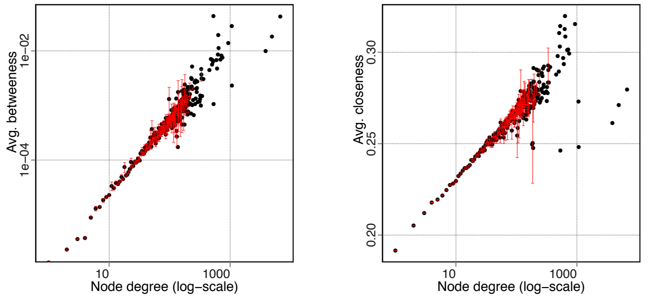
F. Betweenness

The last general topological measures that we evaluate are the *betweenness* and *closeness* centralities. The betweenness centrality of a node is the number of shortest paths between every pair of nodes that passes through the node, and it measures how important the node is in connecting other nodes. The closeness centrality is a measure of how a node is located closely to other nodes. Figure 10 plots the average normalized betweenness and closeness centrality measures against the node degree. As can be seen, there is a strong linear logarithmic correlation between the node degree and these centrality measures, which makes the node degree a viable approximation for the betweenness and closeness indices.

Summary & Implication: In our previous work [34], we have shown that using the node with the highest betweenness centrality as the start or end point in the Maxflow algorithm can improve the reputation accuracy. A problem associated with using this most central node is the high complexity ($O(|V||E|)$) for computing betweenness centrality in unweighted graphs [35]. Although there are approximation algorithms for this measure, but due to the high correlation between the node degree and betweenness centrality, during reputation evaluation process in BarterCast, we can easily use the highest-degree node instead of the most central node.

V. Geographical Characteristics

In this section we consider the nodes in the work-graph from the perspectives of *Autonomous Region* (AR) and *Internet Service Provider* (ISP), and investigate the correlation of the AR and the ISP of neighbor nodes. An autonomous region is a country or a geographical region that according to the IP-to-location mapping is considered as an independent body, e.g., ”Virgin Islands of British”. To obtain the required locality data we use the information collected by *super-peer* nodes in Tribler. When a Tribler client starts, it contacts one of the super-peer machines and gets a set of nodes to contact; the contacted super-peer logs the peer information. Using the



(a) Average betweenness centrality (both axes are in log-scale).

(b) Average closeness centrality (horizontal axis is in log-scale).

Fig. 10. Average betweenness and closeness centrality vs. node degree.

TABLE II
COUNTRY AND ISP LEVEL TRAFFIC INFORMATION.

# AR	intra-AR traffic ratio	# ISPs	intra-ISP traffic ratio
184	0.30	459	0.20

super-peer logged data we are able to find the IP addresses of the peers and locate them at the granularity of ISP. For mapping IPs to locations, we use the IP-to-location service provided by MaxMind[1]. In total, we were able to detect nearly 75% of the peer locations. Using the location information of the peers, we compute the ratio of the data exchanges that happen inside the ARs or ISPs to the total traffic in the network, which are presented in Table II. Besides, we determine the AR and ISP assortativity measures in the work-graph, which show a tendency of the peers to connect to peers in the same AR or ISP. For the work-graph, the AR and ISP assortativity values are 0.0085 and 0.039, respectively.

Summary & Implication: Considering the traffic ratios presented in Table II, it seems that there is a tendency toward having intra-AR and intra-ISP traffic. By further investigation we observed that in a few countries like the USA, the UK, and the Netherlands, the population of Tribler users is so high that statistically encountering a peer in these countries high enough to bias the ratio values. This argument applies to ISP ratio as well, and some huge ISPs cover many peers. Therefore, using the available traffic data we cannot confirm that there is a strong correlation between being in the same AR or ISP and doing a content exchange. Nevertheless, the positive AR and ISP assortativity values indicate a tendency of peers to connect to peers in the same AR or ISP, even though it is not strong.

VI. Peer Behavior and Similarity

In this section, we complement our previous findings about the work-graph by analyzing it from the perspective of the activity of peers. Moreover, using our complementary dataset from the super-peers, we investigate the content-based similarity of neighbor peers in the graph.

In the work-graph the directions and the weights of the edges show the direction and the amount of the content sent and received by a peer. Consequently, the sum of the weights of the outgoing edges of a node shows its contribution to the network, and dividing this value by the total amount of content sent and received gives the *sharing ratio* of the node. Figure 11 shows the CDF of the sharing-ratio values, which vary between the extreme situations of no uploading and only uploading.

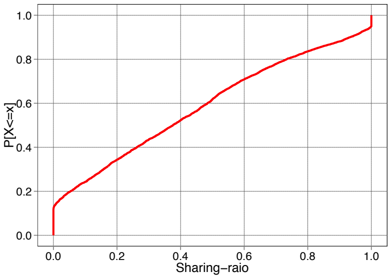


Fig. 11. The CDF of the sharing ratio in the work-graph.

By using a complementary dataset about the activity of peers, which contains the BitTorrent swarms peers have participated in, we can derive the similarity between nodes in the work-graph. These data are logged by the super-peers. Like the locality information, we do not have the whole activity of every peer, but we are able to extract this information for nearly 80% of the nodes.

Based on the set of swarms of each peer and using the Cosine similarity method [36], we study the correlation between having an edge in the graph and participating in a common swarm. To investigate the relation between similarity and having a common edge, we do a number of experiments where we compare the similarity of neighbor nodes in the work-graph with the similarity of neighbor nodes in random isomorphs of this graph. Random isomorphs leave the structure of the work-graph untouched, and nodes have different, but the same number of, neighbors in each isomorph. Like for the original graph, for each isomorph we calculate the similarity of each node to its neighbors and average over each edge. Figure 12 presents the comparison of the CDFs of the similarity values in the original work-graph and the average similarity values obtained through 100 random isomorphs.

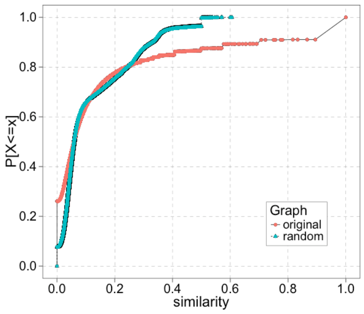


Fig. 12. The empirical CDF of the similarity of neighbor nodes in the work-graph vs. the average similarities in the random isomorphs of the work-graph.

Summary & Implication: From Figure 11 we see that nearly 12% of the nodes are purely passive (sharing-ratio = 0)

[1] http://www.maxmind.com

and do not perform any work for the system, and that nearly 4% are purely active (sharing-ratio = 1). The remaining nodes are divided nearly equally among passive (sharing-ratio < 0.5) and active peers (sharing-ratio > 0.5). Concerning the content-based similarity values, for nearly 80% of the edges the Cosine similarity is roughly equal in the original and the random graphs. For the remaining pairs of neighbors, it is observed that the neighbor similarity in the original graph is higher than the average similarity in the random isomorph graphs. This means that in Tribler, peers have a tendency to connect to similar peers.

VII. Conclusion

In this paper, through studying the BarterCast reputation mechanism from the network science perspective we presented a number of useful insights. In relation to the reputation calculation process in BarterCast, we conclude that if peers apply 5 or 6 hops Maxflow algorithm, they can reach significant portion of the nodes in the network. Besides, instead of using the node with the highest betweenness central value [34], which is expensive to compute, peers can use the highest-degree node as a replacement. Concerning the structure of the network, our measures show that it has a power law degree distribution, a relatively low diameter with strong community structures. Moreover, we observed that there is a positive tendency toward interaction with similar taste peers.

References

[1] M. Meulpolder, J. Pouwelse, D. Epema, and H. Sips, "BarterCast: A practical approach to prevent lazy freeriding in P2P networks," in *Proc. of sixth Int'l Workshop on Hot Topics in P2P Systems, Rome, Italy*, 2009.

[2] J. Pouwelse, P. Garbacki, J. Wang, A. Bakker, J. Yang, A. Iosup, D. Epema, M. Reinders, M. Van Steen, and H. Sips, "Tribler: A social-based peer-to-peer system," *Concurrency and Computation–Practice and Experience*, vol. 20, no. 2, pp. 127–138, 2008.

[3] S. Milgram, "The small world problem," *Psychology today*, vol. 2, no. 1, pp. 60–67, 1967.

[4] M. Granovetter, "The strength of weak ties," *American journal of sociology*, pp. 1360–1380, 1973.

[5] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," *Link Mining: Models, Algorithms, and Applications*, pp. 337–357, 2010.

[6] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 29–42.

[7] C. Wilson, B. Boe, A. Sala, K. Puttaswamy, and B. Zhao, "User interactions in social networks and their implications," in *Proceedings of the 4th ACM European conference on Computer systems*. Acm, 2009, pp. 205–218.

[8] B. Viswanath, A. Mislove, M. Cha, and K. Gummadi, "On the evolution of user interaction in facebook," in *Proceedings of the 2nd ACM workshop on Online social networks*. ACM, 2009, pp. 37–42.

[9] P. Mahadevan, D. Krioukov, M. Fomenkov, X. Dimitropoulos, A. Vahdat *et al.*, "The internet as-level topology: Three data sources and one definitive metric," *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 1, pp. 17–26, 2006.

[10] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *ACM SIGCOMM Computer Communication Review*, vol. 29, no. 4. ACM, 1999, pp. 251–262.

[11] W. Willinger, D. Alderson, J. Doyle, and N. P. S. M. CA., *Mathematics and the internet: A source of enormous confusion and great potential*. Defense Technical Information Center, 2009.

[12] L. Subramanian, V. Padmanabhan, and R. Katz, "Geographic properties of internet routing," in *USENIX Annual Technical Conference*, 2002, pp. 243–259.

[13] S. Kasiviswanathan, S. Eidenbenz, and G. Yan, "Geography-based analysis of the internet infrastructure," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 131–135.

[14] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," *Computer networks*, vol. 33, no. 1, pp. 309–320, 2000.

[15] B. Viswanath, A. Post, K. Gummadi, and A. Mislove, "An analysis of social network-based sybil defenses," in *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 4. ACM, 2010, pp. 363–374.

[16] I. Konstas, V. Stathopoulos, and J. Jose, "On social networks and collaborative recommendation," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 195–202.

[17] S. Kamvar, M. Schlosser, and H. Garcia-Molina, "The eigentrust algorithm for reputation management in p2p networks," in *Proceedings of the 12th international conference on World Wide Web*. ACM, 2003, pp. 640–651.

[18] J. P. R. Delaviz and D. Epema, "Targeted and scalable information dissemination in a distributed reputation mechanism," in *Proceedings of the seventh ACM Workshop on Scalable Trusted Computing (ACM STC)*, ACM. ACM, 2012, pp. 55–66.

[19] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd ed. MIT Press and McGraw-Hill, 2001, pp. 651–664.

[20] L. Li, D. Alderson, J. Doyle, and W. Willinger, "Towards a theory of scale-free graphs: Definition, properties, and implications," *Internet Mathematics*, vol. 2, no. 4, pp. 431–523, 2005.

[21] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 835–844.

[22] A. Clauset, C. Shalizi, and M. Newman, "Power-law distributions in empirical data," *Arxiv preprint arxiv:0706.1062*, 2007.

[23] Q. Vuong, "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica: Journal of the Econometric Society*, pp. 307–333, 1989.

[24] M. Newman, "Mixing patterns in networks," *Physical Review E*, vol. 67, no. 2, p. 026126, 2003.

[25] M. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026113, 2004.

[26] M. Newman, *Networks: an introduction*. Oxford University Press, Inc., 2010.

[27] A. Clauset, M. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, p. 066111, 2004.

[28] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[29] J. Leskovec, K. Lang, and M. Mahoney, "Empirical comparison of algorithms for network community detection," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 631–640.

[30] V. Blondel, J. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, 2008.

[31] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," *Physical Review E*, vol. 74, no. 1, p. 016110, 2006.

[32] U. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Review E*, vol. 76, no. 3, p. 036106, 2007.

[33] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005, pp. 177–187.

[34] R. Delaviz, N. Andrade, and J. Pouwelse, "Improving accuracy and coverage in an internet-deployed reputation mechanism," in *Peer-to-Peer Computing (P2P), 2010 IEEE Tenth International Conference on*. IEEE, 2010, pp. 1–9.

[35] U. Brandes, "A faster algorithm for betweenness centrality*," *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001.

[36] J. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1998, pp. 43–52.