Abstracting .torrent Content Consumption into Two-mode Graphs and their Projection to Content Networks (ConNet)

Andri Lareida, Romana Pernischova, Bruno Bastos Rodrigues, Burkhard Stiller
Communication Systems Group CSG@IfI, University of Zurich
Binzmühlestrasse 14, CH-8050 Zürich, Switzerland
Email: [lareida|rodrigues|stiller]@ifi.uzh.ch, romi@icu.uzh.ch

***Abstract*—Video-on-demand and live streaming services are about to take over video discs. Video streaming services typically cannot compete with the content available in Peer-to-Peer (P2P) file sharing networks. Thus, content providers can use P2P systems to identify content to include in their offer. This work defines a novel method to apply Social Network Analysis (SNA) on video streaming or download traces. Those traces are abstracted int a two-mode graph, which is projected to a content-centric one mode graph (ConNet). SNA measures are used on a ConNet to classify a content-centric graph and provide a general interpretation and insights into the system the traces were collected from. To evaluate the proposed method, real world traces acquired from BitTorrent (BT) swarms sharing movies and television (TV) shows are used to construct 48 hourly graphs to show the evolution of the graph. The results show that the video network can be classified as scale-free, that SNA measures can be used as an alternative popularity indicator, and that the network evolves over time and exhibits diurnal patterns. Finally, this work shows that the proposed method can be applied to real world traces and provides a novel perspective on video consumption.**

I. Introduction

Video content is today's driver of Internet bandwidth consumption not only through streaming but also through peer-to-peer (P2P) file sharing [20]. Streaming services, such as Netflix or Hulu, are constantly updating their catalogs to appeal to the broadest possible customer base to maximize their profits. Recommendations are a key aspect of a video streaming service and critical to maintain and expand a customer base as well as keeping users engaged with the service [2], [9]. It is generally expected that physical media, such as DVDs, will almost disappear in the near future and streaming will take over [17]. Therefore, those recommendation algorithms [9] and catalog management will become even more important.

Social Network Analysis (SNA) traditionally investigates individuals and their social relations, which are modeled as graphs consisting of nodes (individuals) and edges (relations, e.g. friendship) [21]. Modeling a system as a graph enables the use of several SNA measures, which indicate properties of nodes or the whole network, termed node and network centric measures respectively. The main limitation of most SNA measures is that they can only be applied to graphs with one type of node, termed a one-mode graph [21]. Libraries to calculate these measures [1] are well supported and tested and can be easily applied to any one-mode graph. Therefore, SNA methods are not limited to social networks [23]. Furthermore, SNA measures cannot be interpreted or compared to absolute numbers as the measures are depending on the size of a graph.

Therefore, this work proposes a novel method to abstract a one-mode graph from traces comprising: user identification, content identification, time stamp (of view or download). This method consists of a three stage MapReduce job and is therefore applicable to Big Data traces as well as smaller ones. The proposed method is applied to real world BitTorrent (BT) traces as a proof of concept. The resulting graph's edges are analogous to the "watched in the same session" graph used in YouTube recommendations [9]. Since BT is a P2P network the application of SNA to its traces is self-evident. However, the BT network has been thoroughly investigated from a technical perspective [6], [12], [16]. Therefore, using a novel perspective on BT networks, this work applies SNA methods to 200 GB of real world BT traces collected in March 2016 by the VIOLA measurement system [13]. Furthermore, the calculation of SNA measures on the resulting graph provides an example of interpretation of those measures which can be re-applied to other data sources, *e.g.,* traces from video streaming services.

Thus, the contributions of this paper can be summarized as follows:

- A method for abstracting content consumption traces into a two-mode graph and a projection of that graph into a content-centric one-mode graph, termed content network (ConNet).
- A general classification of the content network by comparing relevant SNA measures to those of the three basic network types: random, small world, and scale-free.
- Based on SNA measures calculated from the ConNet an additional measure to identify popular content is provided.
- The evolution of the network over the course of two days is analyzed. Diurnal patterns depending on the filtering level are identified and explained.
- An qualitative interpretation of those measures is given, which can be applied to any graph constructed through the proposed method.

To detail those contributions the remainder of this paper is outlined as follows. Section II introduces the basics of SNA

applied in this work. Section III details the method to abstract the ConNet and presents questions answered by applying the method. Those questions are answered in Section IV together with the results of the ConNet analysis. Finally, the work is concluded in Section V.

II. Background and Related Work

SNA metrics and network models used in this work are briefly explained herein for the reader unfamiliar with SNA. Furthermore, an overview of related work video popularity and SNA in P2P systems is provided in order to highlight the contributions presented in this work.

A. Network Metrics and Models

Social Network Analysis (SNA), provides ways to quantify characteristics of complex networks through measuring them. Measures can be divided into two main groups, node centric and network centric measures. Those measures refer directed and undirected graphs, depending on the type of edges, *e.g.,* an email sent from Alice to Bob (directed) or Alice and Bob sharing an office (undirected). Node centric measures are used to measure point centrality in different ways, providing indicator of influence or power in a network. The node centric measures used in this paper are listed below as defined in [21]:

- **Degree**
- **Closeness**
- **Betweenness**

A network structure can be classified based on a theoretical model. There are three established concepts for network structure as shown in [10], [15]. Random, small-world and scale-free networks differ in various aspects and are characterized based on SNA measures. Each of them implements different real world phenomenons [15], which are skewed degree distribution, clustering, or the small-world effect.

- **Random** networks have been introduced by Erdös and Renyi in 1959 [10], [15]. A random network is constructed by starting with an empty graph having a predefined number of nodes. The predefined number of edges is then placed randomly between these nodes, thus creating a random network. Due to the number of nodes and edges being predefined, random networks of any size and density can be constructed. Because of the random decisions, the degree distribution is a Poisson distribution and therefore the majority of the nodes have a degree close to the mean.
- The **small-world** model was introduced by Watts and Strogatz [15]. It is based on the occurrence of the small-world effect, which states that the average distance between nodes is short. It scales logarithmically with the network size $log(n)$ [15]. This form of closeness is attributable to the existence of individual clusters which are separated from each other and only connected by a few edges. Edges that connect the clusters are therefore long distance and form a short cut between the clusters. In a friendship network this may represent a person that has spent some time overseas and has formed long distance friendships, or politicians and celebrities who tend to travel frequently.
- A **scale-free** network is based on the idea of preferential attachment [10], [15]. Characteristic for this type of network is the preferential attachment property: as a network evolves and more nodes join, there is a probability to which nodes they will connect. Nodes with a high degree have a higher probability of being connected with new nodes than nodes with a low degree. Thus, well connected nodes become even more connected as the network grows. This effect is reflected in the degree distribution: plotting the degree of nodes to the probability of occurring in the network. Therefore, the degree distribution of scale-free networks follows a power law distribution.

Table I summarizes the properties of those three network models presented. Furthermore, the aforementioned models can be distinguished through a comparison with a random network.

B. BitTorrent Measurements

The Gnutella network, which is also a P2P file sharing network, has been analyzed regarding its topology in [18]. Similar to VIOLA [13], a crawler was used by the authors to gain information about the Gnutella network which was then analyzed in terms of degree distribution and clustering. Their findings conclude that Gnutella does not present a pure power law distribution, but benefits from a similar structure nonetheless. These findings provided the base for the analysis of the Internet traffic and its efficiency and scalability. The authors therefore proposed improvements to the protocol that is used by Gnutella based on the conducted research.

SNA is used to analyze various networks, such as the innovators network [4]. The authors focus on the evolution of the network and their findings concluded with the insight that new innovators entered the network close to the core, following the idea of preferential attachment [4]. SNA was also used as a tool to visualize the patent citation network and to gain insight in centrality of popular inventors [22]. They distinguish between weak and strong citations, which depend on the extensiveness of the citation and origin, either from strong rivalry or from colleagues inside of the same company [22]. Also, SNA was used to monitor and analyze social structures and their impact of individuals and teams inside organizations [3]. SNA measures, such as centrality, can be approximated by entropy production which is computationally less intensive [11]. Approximations may be the only option when analyzing big graphs.

A routing scheme for MANETS (Mobile Ad-hoc Networks) using SNA was used to determine the best route in the network [7]. MANETs are self-configuring and can function in the absence of fixed infrastructure. If a node is looking for a destination but does not know the route, it forwards a message to a node that is more central and therefore has a higher probability of knowing the destination. As a metric for centrality the betweenness measure was used. Experimental results showed that the SimBet Routing was as successful as Epidemic routing in delivering messages and even caused less overhead.

TABLE I. RELATION BETWEEN DIFFERENT CLASSES OF NETWORKS.

Class	Average Path Length (L)	Clustering Coefficient (CC)	Degree Distribution
Random	L_R	CC_R	Poisson
Small World	$L_{SW} \leq L_R$	$CC_{SW} \geq CC_R$	≈Poisson
Scale-Free	$L_{SF} \leq L_R$	NA	Power Law

In [19] a measurement study was conducted to collect IP addresses of BT swarms. The crawler described in the paper uses only a single machine to repeatedly query all trackers of a torrent. The torrents were collected from "Mininova" and "The Pirate Bay" torrent portals. The first one included the latest 40,000 torrents from Mininova and was executed 3 times with one week between measurements. The paper states that a complete snapshot took 90 minutes to complete which is a long time span for a snapshot as churn will have some effect during this period. The IP addresses and ISPs are per snapshot. The other two measurements cover fewer torrents but were executed hourly for a full day and the reported numbers are per full day. The analysis of the collected data does not include content-centric abstractions or graph measures.

C. User Behavior in VOD Systems

Understanding user behavior in VOD systems is important for managing such systems, *e.g.,* for resource provisioning or scheduling maintenance windows [24]. [24] analyzed log data of a Chinese VOD streaming service, offering 7,036 videos, over 219 days. The study identified diurnal patterns like they can also be found in BT [13]. Furthermore, content popularity distribution was analyzed, showing a Zipf-like distribution. In this context, popularity was defined as the number of media accesses per video. Additionally, a comparison of session length and video popularity was made, which revealed no correlation between the two.

Another study of user behavior in VOD [5] focused on YouTube and Daum, both offer mainly user generated content. The data was acquired by crawling the websites of the two portals to collect information, such as upload time, video length, views, *etc.*. The data shows a strongly skewed popularity distribution with the top 10% of videos accounting for 80% of the total views. The data also indicates that, generally, older videos are less popular.

[9] gives a simplified explanation of the YouTube recommendation system, having the goal of providing the best recommendations to users to keep them engaged. The core metric used to find candidates for recommendation is co-visitation counts, expressing how often a pair of videos is watched in the same session. The candidate set is expanded by adding also co-visited videos of co-visited videos of the original. Those recommendations are pre-computed in batch processes allowing to include large amounts of data in the computation and serving recommendations in little time. To evaluate their algorithms the team deploys new features to a user group and observes an additional control group to deduce the impact of the new features.

An analysis of this related work presented indicates that there is no known application of SNA methods onto the BT network from a content-centric perspective. However, for other P2P systems, such as Gnutella, SNA methods were used to investigate technical parameters, such as the Gnutella overlay topology [18]. Thus, the ConNet approach developed here is new and the SNA metrics presented promise novel insights into the content consumption behavior of BT users, the identification of content connecting groups of user, the influence of time on such a consumption behavior, or the popularity of content.

III. THE CONNET METHOD

For content providers it is trivial to collect the user data necessary to abstract a ConNet, which requires a tuple consisting of times, user ID (UID), and a content ID (CID), called a record. The method shown in Figure 1 defines the process of abstracting a collection of such records to ConNets and applying SNA to it. First the data transformation from records into a ConNet is explained, followed by the definition of ConNet specific research questions and strategies on how to answer them with SNA measures. The following sub-sections present further details on the abstraction method.

A. ConNet: Mapping the Network

Figure 1 depicts the abstraction method, starting with defining research questions and appropriate time slots, *e.g.,* investigation of day time influence on video consumption in one hour slots. The mapping process and its stages is detailed as a sub process including the outputs of those stages. The processing of the stages suits a MapReduce model and can, thus, scale to large data sets. The details of the individual stages are explained in detail.

Stage 1 transforms a set of tuples into a two-mode graph, depicted in Figure 2. Depending on the data format the records are in, Stage 1 can be left away. However, in the most general case data can contain redundancies, as in the BT case, which can be elegantly removed by a MapReduce job, handling the filtering as well. Filtering is generally necessary since certain time slots will be investigated and compared. If there is more than one time slot the filter can be changed and the whole process repeated for every time slot. Other attributes of records can be used for filtering, *e.g.,* a specific set of users or the location of users if it is available. The MapReduce job abstracts the content providing service users who are connected to a content if they have consumed it, *i.e.,* watched the video. Therefore, the Map phase maps each record to a concatenation of CID and UID. For each combination of CID and UID, the Reduce phase emits an edge into an edge list.

Stage 2 implements the projection of the two-mode graph into a one mode graph, as depicted in Figure 2. The projection removes the user nodes from the two-mode graph and replaces them with fully meshed connections between the contents that user was connected to. Semantically such an edge means “users also watched” similar to the Youtube recommender system [9], an analogous example is the “bought together” used in online

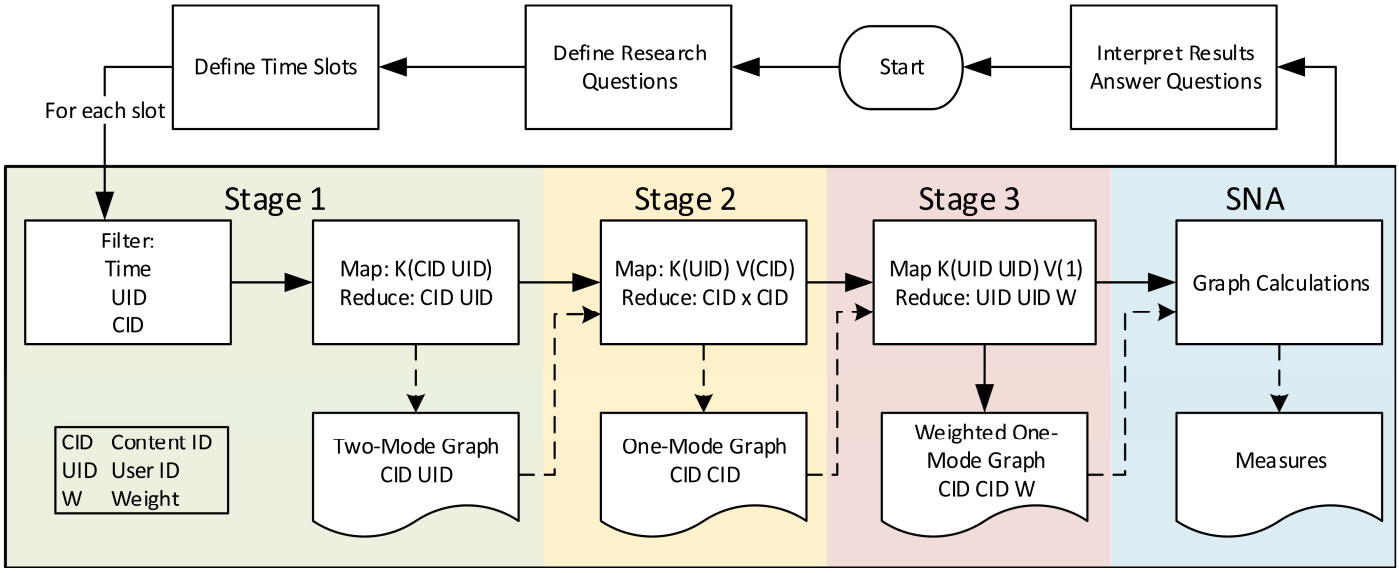


Fig. 1. Diagram of the ConNet abstraction process flow.

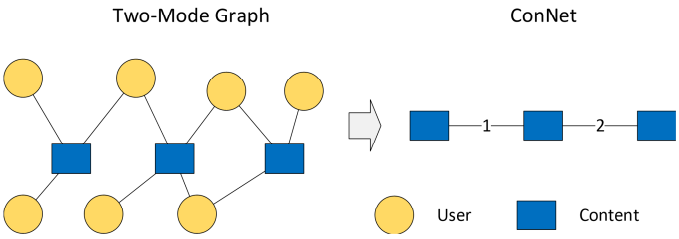


Fig. 2. Abstract BitTorrent two-mode graph and its torrent network projection.

shops. For this purpose the Map phase in Stage 2 maps UIDs to the CID. Thus, the Reduce phase receives a list of CID for every UID and returns an edge for pair of CIDs found, *i.e.,* a permutation of the unique CIDs per UID. The result is an edge list containing an edge for every CID that had a common UID.

Stage 3 reduces the size of the edge list by counting the weight of the edges. Potentially, there are many redundant edges in the edge-list from Stage 2, since users consume multiple contents, nodes will have redundant connections. Thus, another MapReduce job is required which is similar to the word count algorithm. The value 1 is mapped to each edge, the reducer counts all the values for an edge and emits the edge and that count, which corresponds to the weight of that edge. Based on the weighted one-mode graph SNA measures can be calculated. There exist many different such measures, the actual selection of measures to calculate depends on the goals of the analysis. An example of measures and interpretation is given in Section IV.

B. ConNet Characteristics

Based on the content network SNA measures can be calculated independent of where the data is coming from as long as content is linked to users. Thus, the interpretation of those measures and the questions they can answer are similar for all ConNets.

The network classification is done by comparing its characteristics to the characteristics of the three basic network models presented in Section II-A: random, small-world, and scale-free networks. There are three measures that characterize a network: Average Path Length, Clustering Coefficient, and Degree Distribution. Those measures depend on the size of the network. Thus, a random network of the same size, *i.e.,* the same number of nodes and edges, as the ConNet is considered. By calculating the three measures for the network to be classified and the random network allows one to fill in Table I and fit the degree distribution to a power law or a Poisson distribution. Based on this table, a decision on how to classify the network can be taken.

Typically, content popularity is defined by the number of views of a given content, which can be directly calculated from records, not requiring the use of SNA measures. However, for a content provider it might be interesting to use other measures to identify content they should consider for accommodation in their catalogs or content that is important to a small user group. Furthermore, those measures can be used in recommendation systems. The betweenness measure is used to identify content that is relevant for different user groups and therefore has a higher chance to appeal to a bigger audience. Potentially, this content can be cheaper or easier (exclusivity) to license. Besides Betweenness, any node centric measure can be used as indicator for popularity.

Certain ConNets, like P2P file sharing systems, are highly dynamic. Peers are constantly joining and leaving the system and swarms often exhibit a flash crowd behavior [13]. Also new content is constantly added and unpopular content is removed. To visualize the evolution of the network, node and network centric measures demand to be calculated for multiple time slots to illustrate the changes in network structure. For this purpose, two node centric measures, Average Degree and Average Betweenness, and two network centric measures, Average Path Length and Density, are used.

One aspect that needs to be considered when analyzing

BT traces, are peers that are actually executing measurements as in VIOLA [13] or seed-boxes that automatically download new torrents. Since those types of peers are not real users who download movies they like but rather everything that is available, or in case of measurement systems just pretend to download. Therefore, the effects of non-human peers need to be investigated by comparing the full and filtered version of the networks.

IV. The BitTorrent ConNet

This section presents an example of how to apply the abstraction method to real world traces of BT. The details of the transformation are presented first, followed by an analysis of the characteristics of the BT ConNet. Finally, general aspects on the SNA results, their limitations, and the experience with the ConNet abstraction method are discussed.

A. Application of Transformation

Traces from the BitTorrent network come from the VIOLA measuring system [13]. The data set analyzed in this work contains 18'784 torrents from "The Pirate Bay" and "Kick Ass Torrents", which were published in the movie or TV shows categories. The data was collected from March 13 to 14, 2016 amounting to approximately 200 GB. The traces cover 8 Mio. unique IP addresses.

The time slot used is one hour, as it will give 48 data points, making potential diurnal patterns visible. Thus, a ConNet is abstracted for every hour in those two days of data and measures are calculated on each ConNet. The VIOLA data was split in files covering one hour of a day, taking the time filtering out of MapReduce. However, since BT is an open system there are some peers which are used for measurements, such as the VIOLA slaves. Those peers have a high number of contents (torrents) they share. Figure 3 depicts the 100 peers (identified by IP address) which were active in the most torrents over the whole data. The height of the bars indicates the number of torrents. There is a significant drop after the first 11 IP addresses of which 10 stem from the VIOLA measurement slaves, the 11th likely being another measurement system or crawler. The smaller peers represented by the smaller bars are still sharing hundreds of torrents which does not seem to be regular users. Either these are seed-boxes, which automatically download newly published torrents, or Carrier Grade Network Address Translation (CGN) devices, which aggregate many users behind the same public IP address. As investigations have shown, *i.e.,* Figure 4, excluding those top 100 peers changes the ConNet remarkably. However, if the number of maximum torrents shared is lowered to 500 torrents it produces comparable results to lowering the threshold even more. Therefore, herein an unfiltered ConNet (FULL), a ConNet with a torrents per peer maximum of 1000 (MAX1000), and a ConNet with a maximum of 500 (MAX500) are compared. The filter was implemented with a black list used in the first stage's Map phase. To transform the data into a weighted edge list the three stages of the ConNet transformation were implemented accordingly as MapReduce jobs.

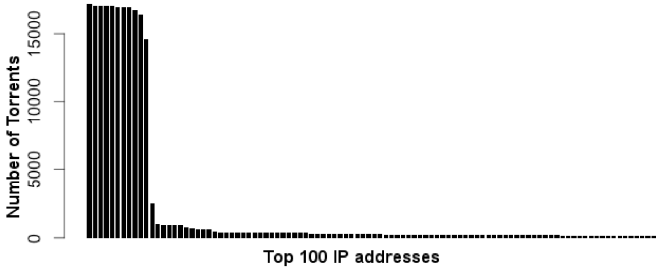


Fig. 3. Torrents shared per IP address.

B. Torrent Network Classification

To classify the torrent ConNet, the Average Path Length (APL), Clustering Coefficient (CC), and Degree Distribution (DD) were compared to random, small world, and scale free graphs. Table II lists the APL for the FULL, MAX500, and MAX1000 including the reduced networks in contrast to random graphs of the same size in the first row. Even the most reduced network has an APL below 2, which lower than in small world graphs, which typically have an average path length of $log(n)$ [15] being ≈ 4 with 16'000 nodes in the this case. Meaning more than 50% of the torrents can be reached in less than two steps. This result is interesting when designing recommendation algorithms using the same type of content network and is consistent with the design choice in the Youtube recommendation system, using two steps to discover content [9].

The second aspect in the comparison is the CC, presented in the second row of Table II. The torrent network has clearly a higher CC than the random graph. This is explained by the high number of edges in the ConNet, causing many connected triplets but few closed triplets. The more peers are excluded, the smaller the CC becomes in both the random and the torrent ConNet, thus, this being accountable to the change in network size. Therefore, it is difficult to draw a conclusion here as the torrent ConNet does not exhibit small world properties and is also very different from a random graph.

The DD presented in Figure 4 shows the full ConNet and the one with top 100 peers excluded. The reduced ConNet fits better to the curve, but the general distribution is very similar to the full ConNet. There is a visible fit to the power law indicated by the straight lines in the graphs. At lower degrees (left side of the graphs) the probability stops to increase exponentially, which is caused by two factors. First, the VIOLA system stopped measuring swarms smaller than 50 peers. Second, earlier studies have found that the popularity of content in BT does not completely follow a power law distribution at the tail [8].

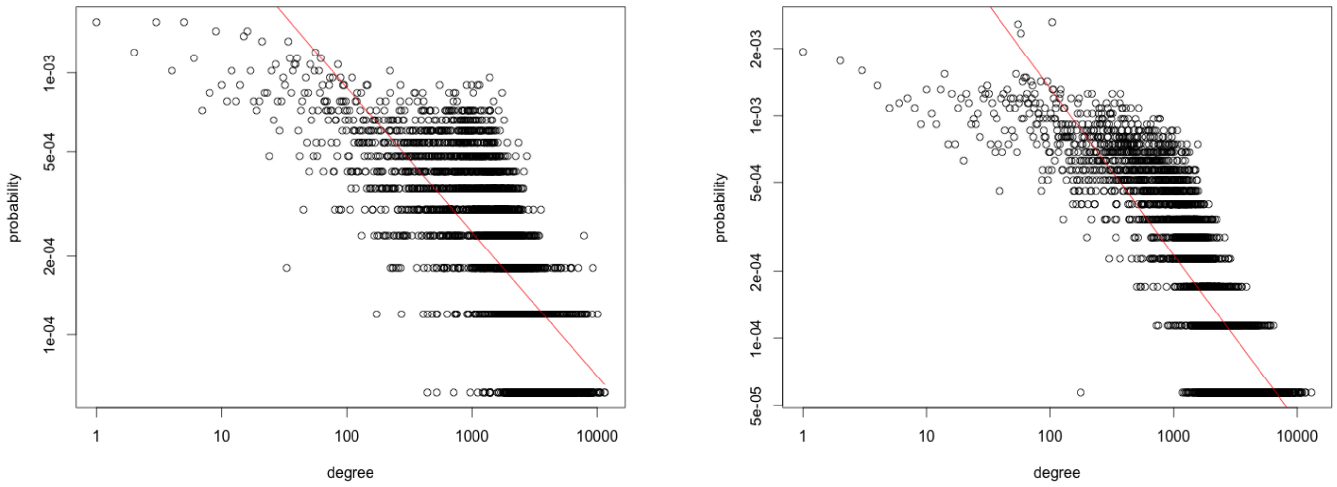
Lastly, the torrent ConNet is classified as a scale free network. It has a clearly non Poisson DD unlike a random network and the low probability for clustering opposes a small world network. Thus, there are a few popular movies or shows, connecting peers from different communities. Thus, it is important to identify those torrents.

C. Popularity of Torrents

The popularity of torrents is typically defined by the number of peers downloading them, analogous to views in a streaming service. Thus, the ranks of degree and downloads

TABLE II. NETWORK CLASSIFICATION MEASURES.

Measure	FULL	FULL Random	MAX1000	MAX1000 Random	MAX500	MAX500 Random
Average Path Length (APL)	1.95	1.91	1.96	1.92	1.98	1.93
Clustering Coefficient (CC)	0.429	0.091	0.361	0.079	0.323	0.066



(a) Full network.

(b) Top 100 IPs excluded.

Fig. 4. TorrentNet degree distribution using the full network and one with the top 100 IPs filtered.

TABLE III. SNA MEASURES ON MARCH 14, 2016 BETWEEN 11:00 AND 12:00 HOURS UTC.

Measure	FULL	MAX1000	MAX500
Number of Nodes	16'940	16'940	16'935
Number of Edges	13'018'406	11'263'700	9'473'873
Density	9.1%	7.9%	6.6%
Average Degree (AD)	1'404	1'203	1'012
Median Degree (MD)	882	882	712
Average Betweenness (AB)	8'008	8'129	8'283
Median Betweenness (MB)	1'710	1'710	1'637
Average Closeness (AC)	0.0000071	0.0000071	0.0000071
Median Closeness (MC)	0.0000071	0.0000071	0.0000071

need be compared to investigate if node degree and popularity are related. The result show that taking the 50 highest degree torrents, 47 of them appear in the top 100 most downloaded torrents. Hence, the degree gives a different perspective on popularity. A degree-based ranking ranks the most-shared torrents highest, reducing the influence of users who just download a single torrent.

Betweenness (BN) tells on how many shortest paths node lies [21], being an indicator for a node's local dependency. If BN is high, *i.e.,* higher than other nodes' B, it indicates a high influence in the network [21] since the particular node lies on many shortest paths that connect groups of nodes. Table III presents the Average and Median Betweenness (AB and MB) in the torrent network. In this network, such a development can be further explained by also considering path length from Table II.

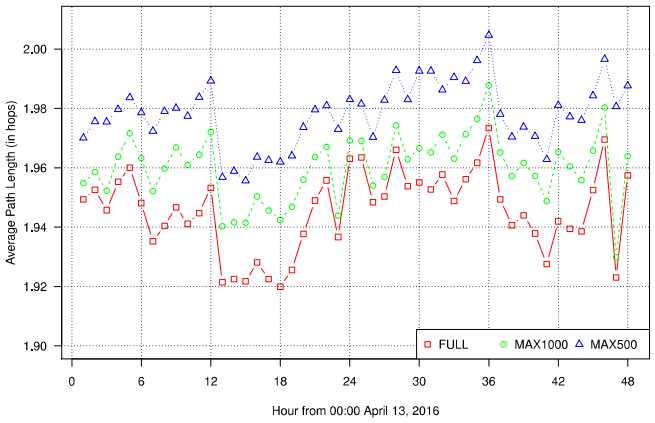
The difference between AB and MB is explained by a power law distribution of BN. Therefore, MB is significantly smaller than AB, which is increased by nodes with extremely high BN (outliers). Those nodes, have a high degree and, thus, lie on many shortest paths, connecting different groups of torrents. These torrents are shared by all types of users as they connect less popular content, *i.e.,* a group, to the network.

D. Torrent Network Evolution

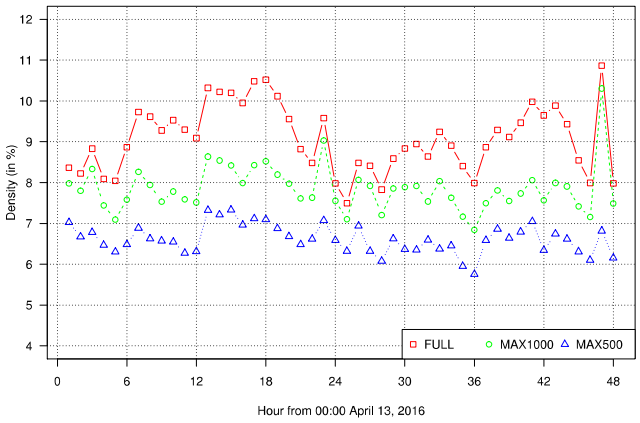
BT is globally popular and, thus, it might not be representative to investigate only one time slot of an hour. The available data set provides records of two days, allowing to investigate the evolution of the torrent ConNet. Figure 5 depicts graphs of four selected SNA measures over the course of those 2 days covered in the data set.

Figure 5a depicts the APL measure over 48 hours. First, the difference between the reduced and the full networks can be observed. The three curves have a similar shape but are shifted upwards with reduction of the network (*i.e.,* MAX500 has higher APL than MAX1000). Second, a diurnal pattern showing a shorter average path length between 10 and 20 hours and a similar trough between 35 and 45 hours can be observed. This is explained by the clean-up task that VIOLA runs daily to remove torrents with swarms older than two days and sizes smaller than 50. This reduces the APL, since those removed torrents are represented as poorly connected nodes in the ConNet. However, those troughs and peaks are on a small scale as the whole graph spans from APL 1.92 hops to APL 2.00 hops.

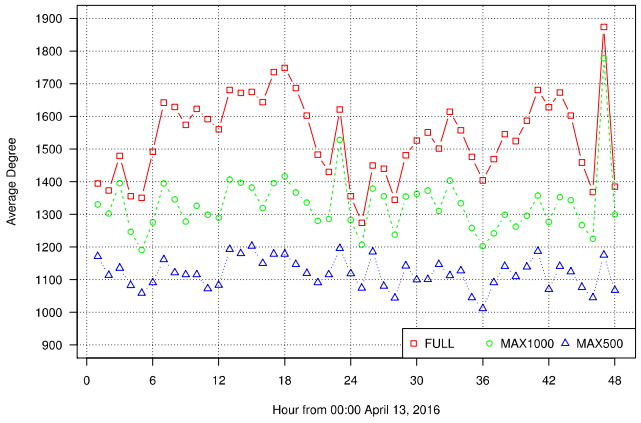
Figure 5b depicts the evolution of density over the two day period. Density gives the relative number of existing connections by the maximum number of potential connections. Therefore, the higher the number of connections, the higher



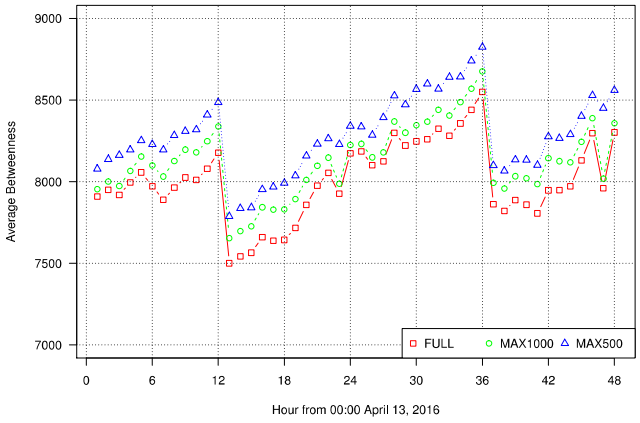
(a) Average Path Length



(b) Density



(c) Average Degree



(d) Average Betweenness

Fig. 5. Evolution of the torrent ConNet over 48 hours of traces.

is the network density, which is inversely proportional to the APL. The density measure shows a diurnal pattern for the FULL and MAX1000 networks but almost nonexistent in the MAX500. Therefore, some nodes introduce many connections to the network and sharing more than 500 torrents. Owing to this effect, the diurnal patterns cannot be accounted to seed-boxes but to CGN which is used when IPv4 addresses are scarce.

Figure 5c shows the mean degree. The curves are similar to the density curves in Figure 5b. Again, the diurnal pattern is weaker in the reduced networks. The reason is the same as for the density and average path length, the CGN which gathers many users behind one public IP address.

Figure 5d shows the Average Betweenness (AB). The small differences between the networks is explained by the lack of certain connections, shortening the paths between many nodes. The average betweenness are constantly increasing until roughly 11:00 hours, followed by a small variance. The increase is due to the addition of new torrents to the measurement, the drop is attributable to a cleanup task. Thus, regarding AB it can be stated that the main influence factor is the number of nodes in the network.

E. Discussion

Table III shows the number of nodes and edges as well as the density. As the nodes represent torrents, one would expect to see 18'784 nodes in the FULL network but there are only 16'940. This observation is explained by the table reflecting only one hour of the 48 hours and some torrents not being shared in this hour. Filtering the network lead to the number of nodes being slightly reduced in the MAX500 network, while the number of edges and, thus, the density showed a greater reduction. This effect is in alignment with the CGN reason since measurement peers and seed-boxes should not influence the content of the BT network. However, with MAX500 regular peers are removed and certain content disappears with them.

The AD decreases with reduction of the network but is always significantly higher than the MD, indicating a power law distribution as shown in Figure 4b. The MD is less affected by filtering. Thus, filtering affects all nodes irrespective of their degrees.

For BN the situation is different. AB is $\approx$4 times higher than MB for all networks. AB increases with the reduction of the network due to the high number paths connecting different torrents, being removed through filtering. At the same time,

MB decreases with filtering due to nodes being removed from the network. Naturally, nodes with low BN are the ones to be removed first.

The Closeness (CN) evaluation was included for completeness, as can be seen from Table III, the difference between the different networks, AC, and MC is in the rounding error of the results. This is an inherent issue of CN as it calculates the inverse of the sum of distances to all nodes. Thus, the more nodes a network has, the smaller is its CN. One solution would be to normalize the sum of distance and use the inverse of the average distance for closeness.

Finally, the application of the abstraction method to a real world data set proves that the method works as the characteristics of the BT ConNet were investigated. In this case the available data set was limited, which is typically not the case for a content providers who store long term records. However, this is a limitation of the data and not of the ConNet method.

V. Summary

This work presented the ConNet method, which connects users to content (*e.g.,* videos) to create a two-mode graph abstraction of a content offering system or service. This two-mode graph is projected into a content-centric one-mode graph, termed ConNet, connecting contents through common users. This method applied to BitTorrent resulted in a torrent ConNet classified as a scale-free network, being resilient and well connected. With an average path length below 2 hops, most content can be easily discovered in 2 hops from any node being a useful input parameter for recommendation systems such as RB-HORST [9], [14].

Network analysis measures can be used as alternatives or supplements to the standard popularity metric of number of views or downloads. Using node centrality measures, content that connects different consumer groups can be identified, providing options to content providers to appeal to new customer groups. Therefore, social network analysis offers new possibilities to investigate and asses content consumption, being relevant for content providing services.

Finally, this work showed how the ConNet method can be applied to real world traces. Therefore, it can be applied to any service or system that can record: user, content, and time records. However, even the BitTorrent traces used herein can be used to learn about the content consumption behavior of its users by analyzing graph measures.

Acknowledgments

This work was supported partially by the FLAMINGO project funded by the EU FP7 Program under Contract No. FP7-2012-ICT-318488.

References

[1] The igraph Core Team, "igraph - The Network Analysis Package," http://igraph.org/, [Online, accessed September 2016].

[2] X. Amatriain, J. Basilico, "Netflix Recommendations: Beyond the 5 Stars," http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html, April 2012, [Online, accessed September 2016].

[3] V. Burger, D. Hock, I. Scholtes, T. Hoßfeld, D. Garcia, M. Seufert, "Social Network Analysis in the Enterprise: Challenges and Opportunities," in *Socioinformatics-The Social Impact of Interactions between Humans and IT*. Springer, 2014, pp. 95–120.

[4] U. Cantner, H. Graf, "The Network of Innovators in Jena: An Application of Social Network Analysis," *Research Policy*, Vol. 35, No. 4, pp. 463–480, May 2006.

[5] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, S. Moon, "I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System," 7th ACM SIGCOMM Conference on Internet Measurement (IMC 2007, San Diego, California, USA, October 2007.

[6] D. Choffnes, F. Bustamante, "Taming the Torrent: A Practical Approach to Reducing Cross-ISP Traffic in Peer-to-Peer Systems," *ACM SIGCOMM Computer Communication Review*, Vol. 38, No. 4, pp. 363–374, October 2008.

[7] E. M. Daly, M. Haahr, "Social Network Analysis for Routing in Disconnected Delay-tolerant Manets," in *Eight ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc'07*. Montreal, Qébec, Canada: ACM, September 2007.

[8] G. Dán, N. Carlsson, "Power-law Revisited: Large Scale Measurement Study of P2P Content Popularity." 9th International Workshop on Peer-to-Peer Systems (IPTPS '10), San Jose, CA, USA, April 2010.

[9] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, D. Sampath, "The YouTube Video Recommendation System," Thr Fourth ACM Conference on Recommender Systems (RecSys 2010), ser. RecSys '10, Barcelona, Spain, September 2010.

[10] O. Hein, M. Schwind, W. König, "Scale-free networks," *Wirtschaftsinformatik*, Vol. 48, No. 4, pp. 267–275, August 2006.

[11] T. Hoßfeld, V. Burger, H. Hinrichsen, M. Hirth, P. Tran-Gia, "On the Computation of Entropy Production in Stationary Social Networks," *Social Network Analysis and Mining*, Vol. 4, No. 1, pp. 1–19, December 2014.

[12] T. Hoßfeld, F. Lehrieder, D. Hock, S. Oechsner, Z. Despotovic, W. Kellerer, M. Michel, "Characterization of BitTorrent Swarms and their Distribution in the Internet," *Computer Networks*, Vol. 55, No. 5, pp. 1197–1215, April 2011.

[13] A. Lareida, S. Schrepfer, T. Bocek, B. Stiller, "Overlay Network Measurements with Distribution Evolution and Geographical Visualization," 2016 IEEE Network Operations and Management Symposium (NOMS 2016), Istanbul, Turkey, April 2016.

[14] A. Lareida, G. Petropoulos, V. Burger, M. Seufert, S. Soursos, B. Stiller, "Augmenting Home Routers for Socially-Aware Traffic Management," 2015 IEEE 40th Conference on Local Computer Networks (LCN), Clearwater Beach, FL, USA, November 2015.

[15] M. E. Newman, "The structure and function of networks," *Computer Physics Communications*, Vol. 147, No. 1, pp. 40–45, August 2002.

[16] J. S. Otto, M. A. Sánchez, D. R. Choffnes, F. E. Bustamante, G. Siganos, "On Blind Mice and the Elephant: Understanding the Network Impact of a Large Distributed System," ACM SIGCOMM 2011, Toronto, Ontario, Canada, 2011.

[17] J. P. Pullen, "5 Reasons Streaming Is Making DVDs Extinct," http://time.com/3921019/streaming-dvds/, June 2015, [Online, accessed September 2016].

[18] M. Ripeanu, I. Foster, A. Iamnitchi, "Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design," *arXiv preprint cs/0209028*, 2002.

[19] R. C. Rumin, N. Laoutaris, X. Yang, G. Siganos, P. Rodriguez, "Deep Diving into BitTorrent Locality," IEEE INFOCOM 2011, Shanghai, China, April 2011.

[20] Sandvine Inc. UCL, "Global Internet Phenomena Report1H," http://www.sandvine.com/downloads/documents/Phenomena_1H_2013/Sandvine_Global_Internet_Phenomena_Report_1H_2013.pdf, [Online, accessed September 2016].

[21] J. Scott, *Social network analysis*. Sage, 2012.

[22] C. Sternitzke, A. Bartkowski, R. Schramm, "Visualizing Patent Statistics by Means of Social Network Analysis Tools," *World Patent Information*, Vol. 30, No. 2, pp. 115–131, 2008.

[23] Valdis Krebs, "Social Network Analysis: An Introduction," http://www.orgnet.com/sna.html, [Online, accessed September 2016].

[24] H. Yu, D. Zheng, B. Y. Zhao, W. Zheng, "Understanding User Behavior in Large-scale Video-on-demand Systems," *SIGOPS Oper. Syst. Rev.*, Vol. 40, No. 4, pp. 333–344, Apr. 2006. [Online]. Available: http://doi.acm.org/10.1145/1218063.1217968