How community-like is the structure of synthetically generated graphs?

Arnau Prat-Pérez
DAMA-UPC
Universitat Politècnica de Catalunya
aprat@ac.upc.edu

David Dominguez-Sal
Sparsity Technologies
david@sparsity-technologies.com

ABSTRACT

Social-like graph generators have become an indispensable tool when designing proper evaluation methodologies for social graph applications, algorithms and systems. Existing synthetic generators have been designed to produce data with characteristics similar to those found in real graphs, such as power-law degree distributions, a large clustering coefficient or a small diameter. However, real social networks are organized into higher level structures, called communities, that are not explicitly considered by these generators. In this paper, we study the statistical features of the community structure found in real social networks, and compare them to those generated by the LFR and LDBC-DG generators. We found that communities show multimodal features, and thus are hard to generate with simple community models. According to our results LDBC-DG draws realistic community distributions, even reproducing the multimodality observed.

1. INTRODUCTION

Real data from social networks is generally difficult to collect and distribute. Collecting data usually requires dedicated Internet connections with high bandwidth, and large amounts of time to obtain meaningful datasets. Furthermore, social data is created by real people, which produces many privacy issues that limit the data collection procedures. Finally, once a real dataset has been collected, it has a fixed data size, which may not fit the specific needs of the applications consuming it. Making smaller subsets of the graph by sampling should be done carefully or biases may appear. On the other hand, extrapolating to larger graphs is even more difficult since the graph structure must be accurately modeled. Last but not least, the distribution of large datasets usually implies hard disk shipping or dedicated servers in the Internet. Under these circumstances, synthetic data generators have become an extremely useful tool to test applications on scenarios that simulate realistic data without the previously described problems.

Proceedings of the Second International Workshop on Graph Data Management Experience and Systems (GRADES 2014), June 22, 2014, Snowbird, Utah, USA.

One of the most prominent families of graph algorithms, which have acquired a significant relevance during last years, is community detection algorithms. Communities are conceptually defined as groups of nodes that are more structurally connected among them than with the rest of the graph. They indicate sets of users with similar interests or profiles. Some of the existing graph generators, such as RMat [1] and MAG [4], were designed to mimic the large scale structure of the network, such as the degree distribution following a long tail power law, a large clustering coefficient and a low diameter. However, metadata about the organization of nodes as communities is not produced by these generators. A different approach is performed by LFR [5], which defines communities as sets of nodes following certain data models and builds the whole network according to these distributions. However, it has not been studied the degree of correlation with real datasets. On the other hand, generators such as the LDBC data generator [3] (LDBC-DG), which simulates a real social network such as Facebook, provide metadata that can be used to enumerate the communities. Graph generators that create community structures, and specially if they explicitly output them, is fundamental for the evaluation of community detection algorithms as well as designing robust benchmarks for graph applications.

There are some works analyzing the quality of the communities output by synthetic generators. In [6], Orman et al. study the aggregated characteristics of the communities generated by LFR. They conclude that LFR is able to reproduce some of the characteristics found in real networks, such as the community and degree distributions, the degree of transitivity (clustering coefficient) and degree correlations, at a network level. However, these last two properties are severely affected by the mixing factor parameter, and become unrealistic when it is above 0.5 [6]. On the other hand, in [2] the authors study the density of the ground truth communities found in real data compared to that expected in a random graph with the same degree distribution. They find that ground communities are clearly denser, which agrees with the informal definition of a community.

The main goal of this paper is to analyze the structure of communities in real graphs, and compare this to those communities output by existing graph generators, such as the LFR and the LDBC-DG generators. Compared to existing work where the analysis is performed at a network level, we focus our study on structural properties of the community and its level of isolation from the graph. We use the following indicators: clustering coefficient, triangle participation

	Vertices	Edges	Communities
Amazon	334,863	925,872	151,037
Dblp	317,080	1,049,866	13,477
Youtube	1,134,890	2,987,624	8,385
LiveJournal	3,997,962	34,681,189	287,512

Table 1: Characteristics of the test graphs.

ratio, bridges, diameter, conductance and size. With this objective in mind, we first compare the structure of ground truth communities from different graph sources to quantify their degree of similarity. Then, we compare those similar patterns, to the communities found by the generators.

Our study reveals three important findings. First, we observe that the communities found among different graphs follow quite similar distributions. Second, we observe multimodal distributions for several indicators when we study the distribution of the communities in a graph. This shows that communities in a single graph have diverse nature, and are difficult to fit with a single model. Third, we see that the LDBC-DG is able to mimic more characteristics of real graphs than LFR.

The rest of this paper is structured as follows. In Section 2, we describe the datasets with ground truth communities. In Section 3, we describe the structural indicators. In Section 4, we describe the synthetic graph generators. In Section 5, we describe the evaluation methodology. In Section 6, we explain the results obtained; and in Section 7, we conclude the paper.

2. DATASETS WITH GROUND TRUTH

We analyze several real graphs with annotated communities as ground truth from the SNAP repository[1]. The communities were annotated automatically according to a procedure that relies on the metadata available in those networks [8]. Such datasets are typically used in state of the art evaluations of community detection algorithms [9]. Given their automatic annotation procedures and the large scale of the datasets, some communities may include noise, but it is out of the scope of this paper to discuss the precision of the annotation method. Given the homogeneity in data distributions among graphs that we detected in Section 6, we can conclude that this noise is regular enough to shape defined data distributions.

The datasets come from different data sources and have varied semantics. Table 1 summarizes the characteristics of the graphs and the interpretation of the communities in the graphs is as follows:

Amazon: This graph represents a network of products, where each vertex is a product and an edge exists between two products if they have been co-purchased frequently. A community is a connected component of the subgraph of products that belong to a product category in Amazon.

DBLP: This graph represents a network of coauthorships, where each vertex is an author and two authors are connected if they have written a paper together. A community is a connected component of the subgraph of authors that published in a conference or a journal.

[1] `http://snap.stanford.edu`

Structural Indicator	Definition
Clustering Coefficient	$\frac{3 \cdot t(S)}{\sum_{x \in S} \lvert N(x) \cap S \rvert \cdot (\lvert N(x) \cap S \rvert - 1)}$
Triangle Participation Ratio (TPR)	$\frac{\lvert \{x \in S : t(x,S) > 0\} \rvert}{\lvert S \rvert}$
Bridge ratio	$\frac{2 \cdot bridges(S)}{\sum_{x \in S} \lvert N(x) \cap S \rvert}$
Diameter	$\frac{diameter(S)}{log(\lvert S \rvert) + 1}$
Conductance	$\frac{\sum_{x \in S} \lvert N(x) \cap (G \setminus S) \rvert}{\sum_{x \in S} \lvert N(x) \rvert}$

Table 2: Structural indicators.

Youtube: This graph represents the Youtube social network, where each vertex is a user and two users are linked if they have established a friendship relation. A community is a connected component of the subgraph of users that belong to a group.

LiveJournal: This graph represents the Livejournal social network Similar to the Youtube network, the vertices are the users, which establish friendship relationships with other users. A community is a connected component of the subgraph of users that belong to a group.

3. STRUCTURAL INDICATORS

For each of the real graphs, we analyze the characteristics of their ground truth communities by means of a set of structural indicators, which help us create a profile of the community structure in these real networks. Before describing the structural indicators we introduce some notation. Let $G(V, E)$ be a graph with a set of vertices V and a set of edges E. Let $N(x)$ be the set of neighbors of x in the graph. Let $t(x, S)$ be the number of triangles that vertex x closes with and only with vertices in set S, and $t(S)$ be the number of closed triangles in the subgraph induced by S. Table 2 shows the structural indicators used and their definitions. On the one hand, Clustering Coefficient, Triangle Participation Ratio (TPR), Bridge[2] ratio and Diameter have been selected in order to get an insight of the internal structure of the communities, that is, we quantify how well and with which structure are the members of the communities interconnected. On the other hand, we select conductance to measure the level of isolation of the communities with respect to the rest of the graph. Finally, we also consider the size as an extra indicator of the communities' profile found.

4. SYNTHETIC GRAPH GENERATORS

We select LFR and the LDBC-DG, because they provide output of the communities created in the dataset. LFR outputs community identifiers with its corresponding members, and the LDBC data generator creates groups of users that we consider as communities:

LFR: LFR was designed as a benchmark for evaluating community detection algorithms [5]. Compared to other graph

[2] A bridge is an edge whose removal disconnects the induced subgraph of the community

generators, its principal characteristic is that the building procedure is based on creating a graph that connects communities. LFR starts by generating a set of communities of different sizes following a power law distribution. Then, the edges between the vertices in the graph are created in such a way, that they follow power law distributions and for each vertex the mixing factor is fulfilled. The mixing factor is a parameter indicating the percentages of neighbors of every vertex that belong to a different community than that the vertex belongs to. A recent study indicated that communities are too well defined, and do not capture the noise found in real data [6]. We downloaded the generator from the author's website.

LDBC-DG: The LDBC-DG is a fork of the S3G2 graph generator [7], which is customized to build social network datasets, which is used in the LDBC benchmarking initiative [3]. The LDBC aims at designing realistic and meaningful benchmarks for linked database systems, namely RDF and graph databases. The LDBC generator generates complex synthetic social-networks with many attributes related to the users and its activities in the network. The resulting schema is similar to the contents available in Facebook. For example, users have attributes that indicate their personal description (name, born place, school/university, etc); the friends of a user; posts and photographs created by a user; groups created by users indicating interests... It starts by generating a set of users with attributes following distributions found in the real world. Then, they sort the users in successive Hadoop jobs by different correlated attributes (i.e. user interests, user universities, etc...) and create friendships between users using a sliding window procedure, where users close in the window have a higher probability to be friends. Following this schema, we create the communities using a similar procedure to the one described in the ground truth, by setting as a community each connected component of users (using the friends relation) that belong to a group. We downloaded the latest available version of LDBC-DG from the Github repository of LDBC on 30th March 2014.

5. EXPERIMENTAL SETUP

Synthetic graph generators have several parameters that can be tuned to produce graphs of different characteristics. For both LFR and LDBC-DG, we generate a network with 150K users. In the case of LFR, we set the average and maximum degree to 10 and 400 respectively, and the minimum and maximum community size to 10 and 10000, respectively. One third of the nodes are set as overlapping nodes, and belong to three different communities instead of one. All these parameters have been set up as to mimic the characteristics found in the ground truth data. Finally, we have set the mixing factor of LFR from 0.1 to 0.5, therefore generating five networks named LFR1, LFR2, LFR3, LFR4 and LFR5, which in this range is expected to generate networks with communities [5]. For the LDBC-DG, we generate a single network using the default LDBC-DG parameters, which are fit to real data. The rest of the parameters for both generators are set to the default values, which are reported to generate realistic social network distributions [3, 5].

For each community, we compute all six structural indicators. Then, we analyze each indicator individually. We take each community as a sample and draw a histogram distribution. Then, we study the correlation between all pairs

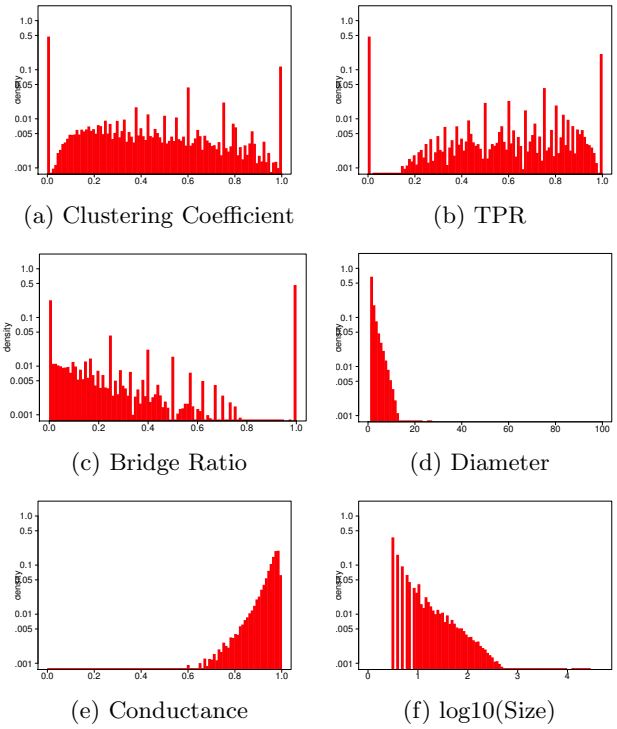


Figure 1: Distribution of the statistical indicators for the Livejournal graph.

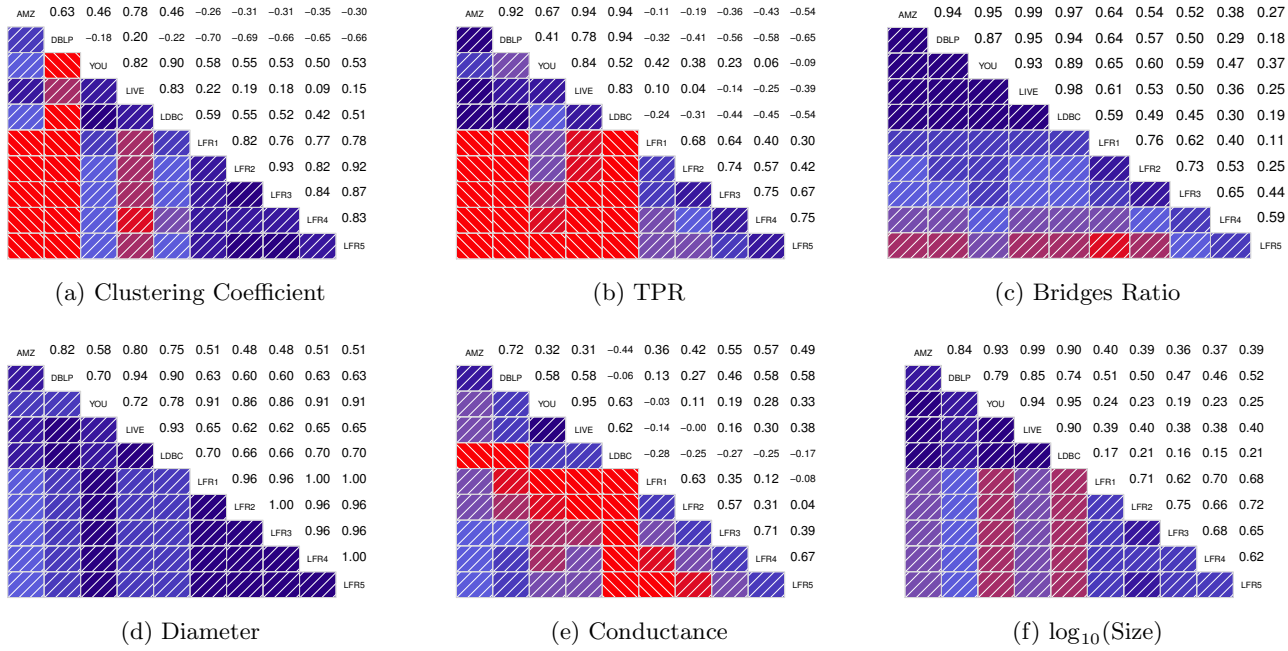
of histograms, by computing the Spearman correlation rank for each pair of graphs. The Spearman correlation rank tests is a non parametric test that quantifies if two variables are monotonically related.

6. RESULTS DISCUSSION

We divide the results' discussion into three parts. We start by comparing the distributions of the four real graphs among them and study their variance. Second, we review the community structure output by the LDBC-DG compared to that found in real graphs. Finally, we do the same for the LFR benchmark.

Real Graphs: Figure 1 shows the distributions of the statistical indicators for the Livejournal graph. We take the Livejournal graph as a representative of the rest of the graphs, which are reported in the Appendix. For the rest of the real graphs, the distributions show similar characteristics as shown by the Spearman correlation tests of Figure 2.

We start by analyzing the internal structure of Livejournal communities, hence we focus our attention on the clustering coefficient, the bridge ratio, the TPR and the diameter. Figure 1(a) shows a multimodal distribution. The largest peak contains 44% of the ground truth communities, with a clustering coefficient between 0 and 0.01. This indicates that many communities have a small percentage of closed triangles. But, when we looked into detail, we found that many of those communities without triangles were very small and lots had only three vertices (59% of them). The second largest peak are communities with a clustering coefficient between 0.99 and 1, which are quasi-cliques or cliques and contain 11% of the communities. The rest of the commu-



(a) Clustering Coefficient (b) TPR (c) Bridges Ratio

(d) Diameter (e) Conductance (f) $\log_{10}$(Size)

Figure 2: Spearman rank correlation coefficient of the distributions between the different communities and structural indicators.

nities fall into intermediate ranges. A similar multimodal result is seen for the TPR and the bridge ratio (Figures 1(b) and (c) respectively) with the two peaks at the extremes and with a trend towards participating in triangles and not having bridges in the central modal group. This multimodal distribution suggests that communities are not an homogeneous entity that can be described with a single model.

In Figure 1(d), we see that the bulk of the communities has a small diameter: 84% have a diameter smaller than five. This is because ground truth communities are well connected and small in many cases. We observe that conductance tends to be high and thus communities are not very well isolated as depicted in Figure 1(e). If we look at Figure 1(f), most of communities (about 74%) have a size smaller than 10. In the last three subfigures we observe that the largest fraction of the communities is small, have very small diameters, and are not very well isolated. For the three indicators, we observe a power law decay towards communities that depart from the typical community.

Figure 2 shows the correlograms of the Spearman rank correlation coefficient between the distribution of the different structural indicators for each pair of graphs. The upper half of the matrix shows the numerical score given a pair of variables. On the other hand, the lower half shows a color gradient, where two variables are correlated if they approach dark blue, while they are not correlated (or inversely correlated if negative) if they approach red.

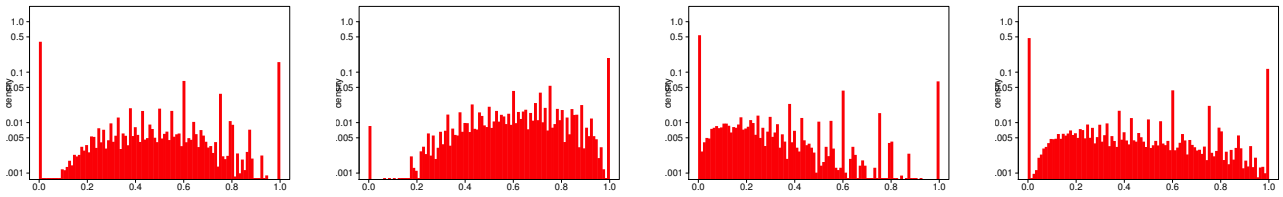
The first four entries correspond to the real graphs. Broadly speaking, we observe that all four graphs show similar patterns for the six indicators. The correlation is specially strong for the bridge ratio, where the rank is over 0.9 for most of pairs of real graphs. The diameter, size and TPR distributions also show important correlations.

We observe that the less correlated distributions are for clustering coefficient and conductance, although correlation is still present. The correlations shown in Figure 2(a) indicate that there are differences between the clustering coefficient distributions for the real graphs, which can be visually compared in Figure 3. First, Youtube and Livejournal have a similar distribution, slightly biased to the left, and having similar peaks at their extremes. Second, Dblp is the graph with a distribution more biased to highly clustered communities. Furthermore, the peak extremes of the Dblp distribution are inverted compared to the rest. This explains why Dblp is not correlated with Livejournal and Youtube. Finally, Amazon lies between Dblp and Livejournal with a more centered distribution.

We see in Figure 2(e) that graphs have two types of conductance distributions. Figure 4 depicts that the conductance distribution of Amazon and Dblp is more diverse (for conductance, the smaller the better). Specifically, 63% and 73% out of the total number of communities for the Amazon and Dblp graphs respectively, have a conductance larger than 0.5. For Youtube and Livejournal, the distribution is more skewed towards the right of the chart and 98% and 99% of the communities have a conductance larger than 0.5.

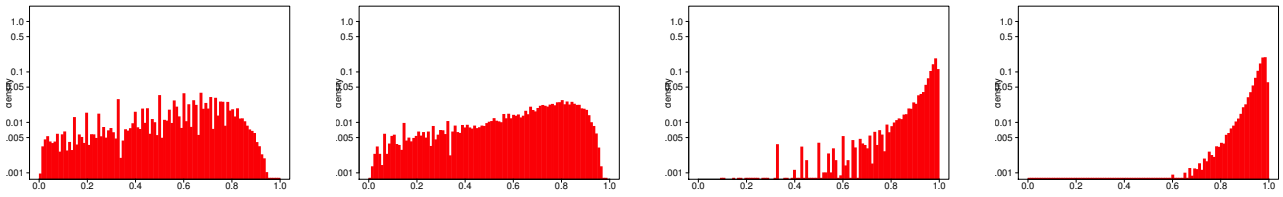
LDBC-DG Graph: Figure 5 shows the distributions of the structural indicators for the LDBC-DG graph and the fifth row in Figure 2 shows the correlation of each plot with the real datasets. We observe that for most indicators the synthetic distributions are considerably similar to those for the real graphs, specially for Youtube and Livejounal.

First of all, the LDBC-DG reproduces the multimodal distributions of the clustering coefficient, the TPR and bridge ratio (Figures 5(a-c)). The multimodal clustering coefficient distribution of LDBC-DG shows a central part biased towards communities with a small clustering coefficient. This is similar to what we see for Youtube and Livejournal



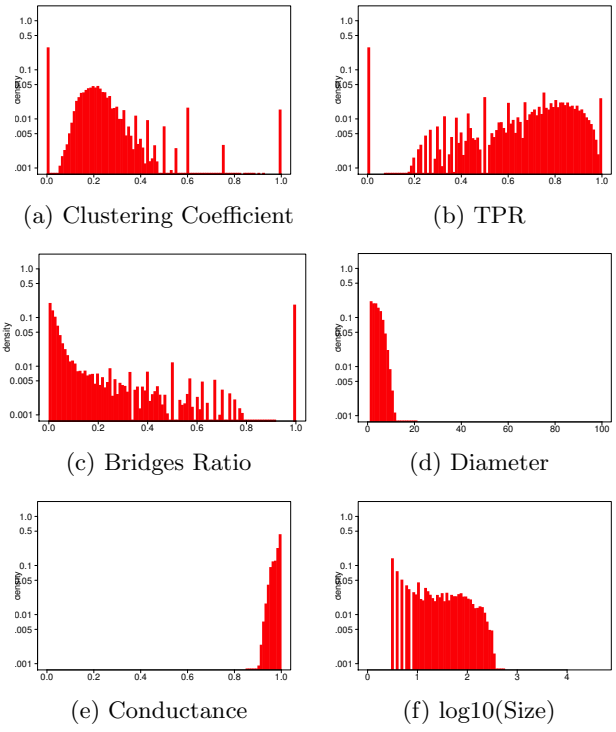
(a) Amazon (b) Dblp (c) Youtube (d) Livejournal

Figure 3: Clustering coefficient distribution of real graphs.



(a) Amazon (b) Dblp (c) Youtube (d) Livejournal

Figure 4: Conductance distribution of real graphs.



(a) Clustering Coefficient (b) TPR

(c) Bridges Ratio (d) Diameter

(e) Conductance (f) log10(Size)

Figure 5: Distribution of the indicators for the LDBC-DG graph.

graphs. We find that the generator distributes evenly the triangles among the members of the communities, as shown by TPR in Figure 5(b). More specifically, 63% out of the total number of communities have a TPR larger equal or larger than 0.5. Figure 5(d) shows the diameter distribution of the LDBC-DG communities. Compared to those found in real graphs, LDBC-DG communities have a slightly larger diameter, with 71% out of the total number of communities with a diameter less than 6.

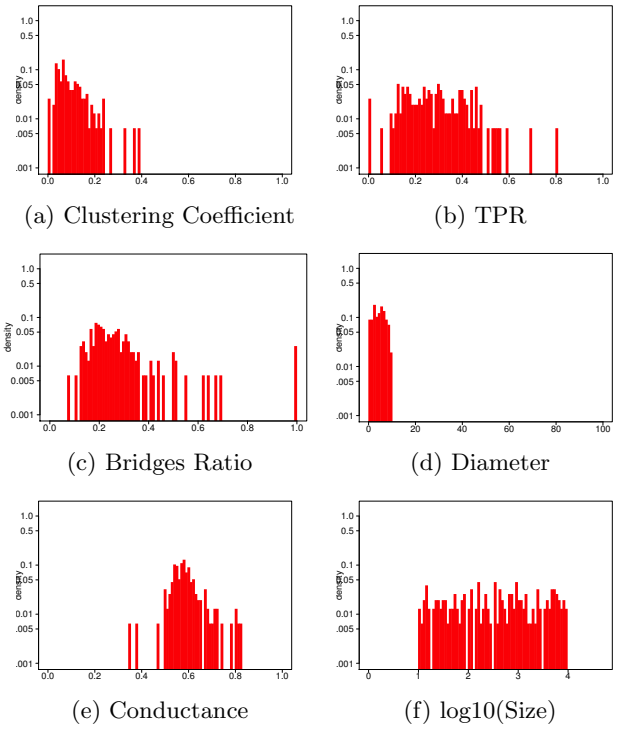
When we turn to analyze the conductance, as shown in Figure 5(e), we see that as with the real graphs, the LDBC-DG communities tend to have a large conductance, similarly to those found for Youtube and Livejournal. However, we note that the distribution is significantly more skewed to the right Thus, LDBC-DG communities are less well isolated than those in the real datasets.

In general, we see that the LDBC-DG reproduces many of the characteristics found in real graphs, specially those found in Youtube and Livejournal. Since LDBC-DG models an online social network data, it seems natural that the communities generated resemble more the datasets from online social networks than the product and coworker network.

LFR Graphs: For the LFR graph, only the diameter distribution shows a strong correlation to those found in real graphs as shown in the last five rows of Figure 2. For the rest of the indicators, the degree of correlation is moderate or weak, though it varies depending on the mixing factor configuration. In order to better understand the characteristics of the community structure of the graphs output by the LFR generator, we show the distributions for the mixing factor 0.3 configuration in Figure 6.

In contrast to LDBC-DG, LFR does not produce the multimodal distribution for clustering coefficient (Figure 6(a)) observed in real graphs. LFR does not produce communities with a large clustering coefficient. According to Figure 6(b), the TPR distribution also lacks a peak for large participation ratios, and in contrast to LDBC-DG it also lacks the peak for the low TPR modality found in real graphs.

The bridge ratio (Figure 6(c)) distribution of LFR has moderate correlation to the real data, but the peak on the left extreme is missing and the peak on the right is smaller than the real ones. Diameter distribution is quite similar to the one found for the real data, but with some more large diameter communities.



(a) Clustering Coefficient (b) TPR

(c) Bridges Ratio (d) Diameter

(e) Conductance (f) log10(Size)

Figure 6: Distribution of the indicators for the LFR3 graph.

Conductance performs a poor match with the real datasets. LFR produces a distribution centered in a certain value of conductance, as shown in Figure 6(e). This peak depends on the mixing factor (see Appendixes for more details), and goes towards the left when the mixing factor is large. Then, configurations of LFR with larger mixing factors produce more realistic conductance plots because they have larger conductances. However, these larger mixing factors, such as LFR5, are much worse in terms TPR, bridge ratio and size as seen in Figure 2.

We have observed that the main weakness of LFR is the regularity of the communities created. Since all the communities follow a single model, LFR is not able to create the multimodal distributions present for some indicators.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we have seen that the distribution of the structure indicators of the communities are very homogeneous *among* datasets. This shows that communities are present in many environments, with similar traits. The most distinctive factor is conductance, which in our datasets shows moderate differences between social networks and other network sources. Given the small number of networks analyzed, it would be interesting to check as future work if this difference can be attributed to the data origin or simply to a larger variance in the distribution.

Given a network, we also observe that the indicators for *each* dataset related to clustering coefficient and presence of the bridges show multimodal distributions. This suggests that communities do not fit a single model, but a mixture of several statistical distributions. Therefore, the creation of realistic communities (and also algorithms to detect them) should consider this diversity and apply different models to generate this community variety.

We observed that LDBC-DG produces communities more similar to those found in the social networks than LFR. Our more detailed analysis shows that the LDBC-DG produces community distributions similar in some aspects to those found in real graphs, specially Youtube and Livejournal, which are social networks. The generator produces communities with a low clustering coefficient, which includes communities in both the small, intermediate and high range of TPR. It is also able to partially reproduce the multimodality observed in the real data where communities with peaks for very large clustering coefficient, TPR and bridge ratio are observed. Regarding to the community size and diameter, LDBC-DG produces larger and broader communities than those observed in real data. We also observe that communities are more similar in terms of isolation to other social networks than other types of networks.

8. ACKNOWLEDGEMENTS

Arnau Prat-Perez thanks the Ministry of Science and Innovation of Spain and Generalitat de Catalunya, for grant numbers TIN2009-14560-C03-03 and GRC-1187 respectively. Arnau Prat-Perez thanks the EU FP7 project LDBC (FP7-ICT2011-8-317548) for funding the LDBC project. David Dominguez-Sal thanks the Ministry of Science and Innovation of Spain for the grant Torres Quevedo PTQ-11-04970.

9. REFERENCES

[1] Deepayan Chakrabarti, Yiping Zhan, and Christos Faloutsos. R-mat: A recursive model for graph mining. In *SDM*, volume 4, pages 442–446. SIAM, 2004.

[2] Marek Ciglan, Michal Laclavik, and Kjetil Nørvåg. On community detection in real-world networks and the importance of degree assortativity. In *KDD*, pages 1007–1015, 2013.

[3] Renzo et al. The linked data benchmark council: a graph and rdf industry benchmarking effort. In *To appear in SIGMOD Record*. ACM.

[4] Myunghwan Kim and Jure Leskovec. Multiplicative attribute graph model of real-world networks. *Internet Mathematics*, 8(1-2):113–160, 2012.

[5] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Phy. Rev. E*, 78(4):046110, 2008.

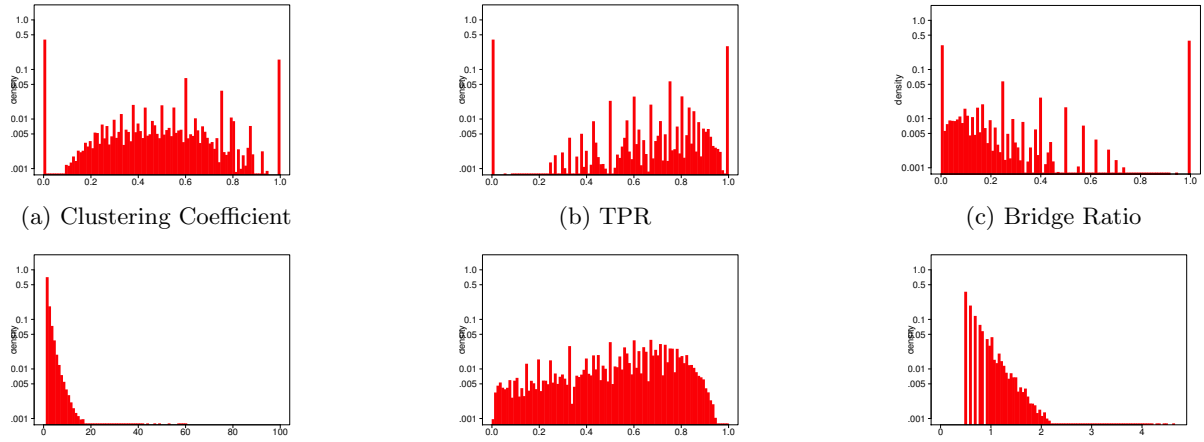
[6] Günce Keziban Orman and Vincent Labatut. A comparison of community detection algorithms on artificial networks. In *Discovery Science*, pages 242–256. Springer, 2009.

[7] Minh-Duc Pham, Peter Boncz, and Orri Erling. S3g2: A scalable structure-correlated social graph generator. In *TPCTC*, pages 156–172. Springer, 2012.

[8] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. In *MDS*, page 3. ACM, 2012.

[9] Jaewon Yang and Jure Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *WSDM*, pages 587–596. ACM, 2013.

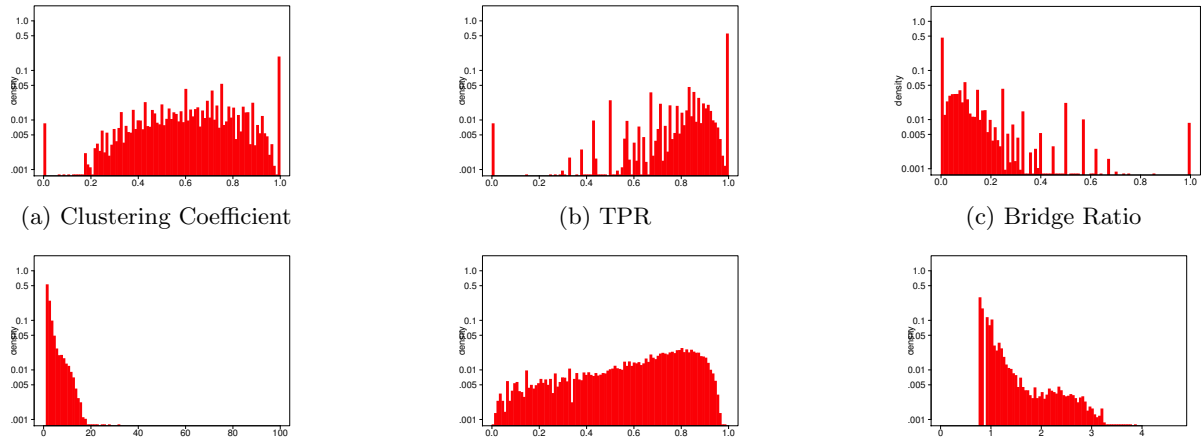
APPENDIX



(a) Clustering Coefficient (b) TPR (c) Bridge Ratio

(d) Diameter (e) Conductance (f) log10(Size)

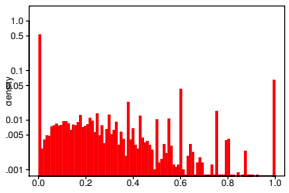
Figure 7: Distribution of the statistical indicators for the Amazon graph.



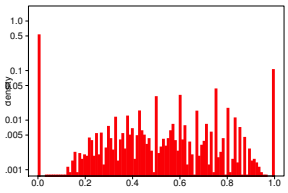
(a) Clustering Coefficient (b) TPR (c) Bridge Ratio

(d) Diameter (e) Conductance (f) log10(Size)

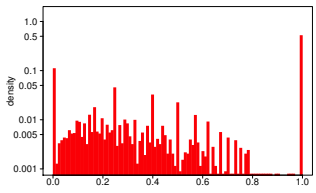
Figure 8: Distribution of the statistical indicators for the Dblp graph.



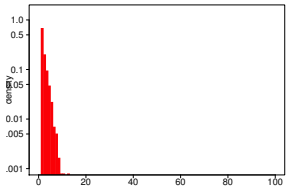
(a) Clustering Coefficient



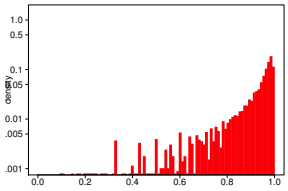
(b) TPR



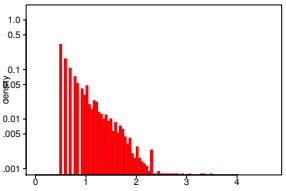
(c) Bridge Ratio



(d) Diameter

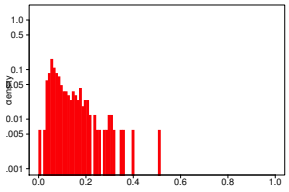


(e) Conductance

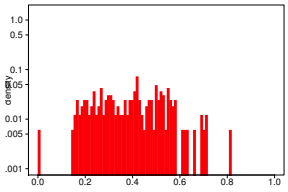


(f) log10(Size)

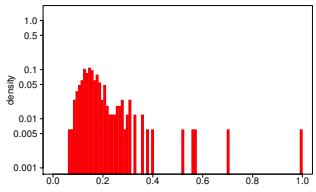
Figure 9: Distribution of the statistical indicators for the Youtube graph.



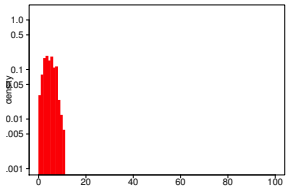
(a) Clustering Coefficient



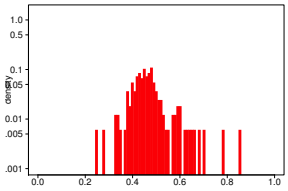
(b) TPR



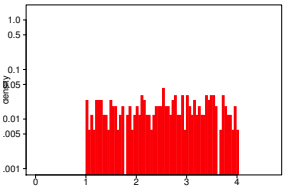
(c) Bridge Ratio



(d) Diameter

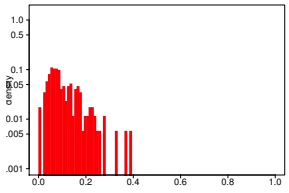


(e) Conductance

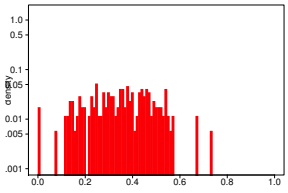


(f) log10(Size)

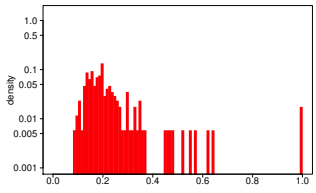
Figure 10: Distribution of the statistical indicators for the LFR1 graph.



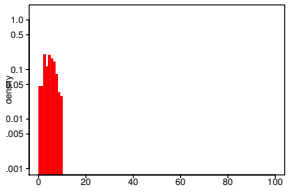
(a) Clustering Coefficient



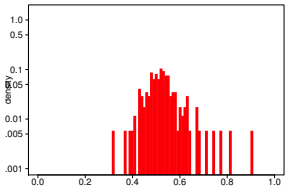
(b) TPR



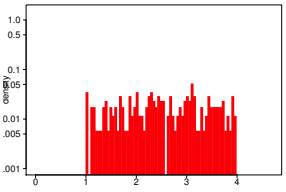
(c) Bridge Ratio



(d) Diameter

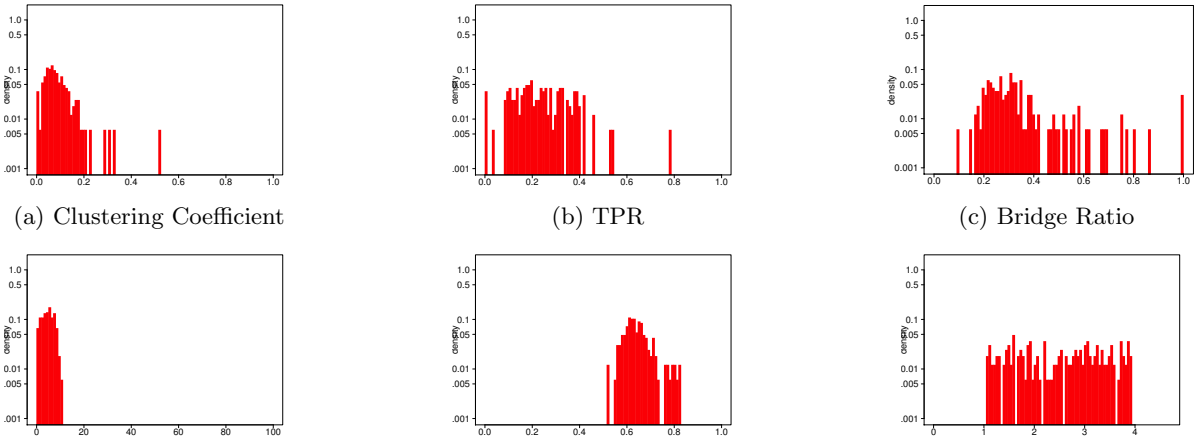


(e) Conductance



(f) log10(Size)

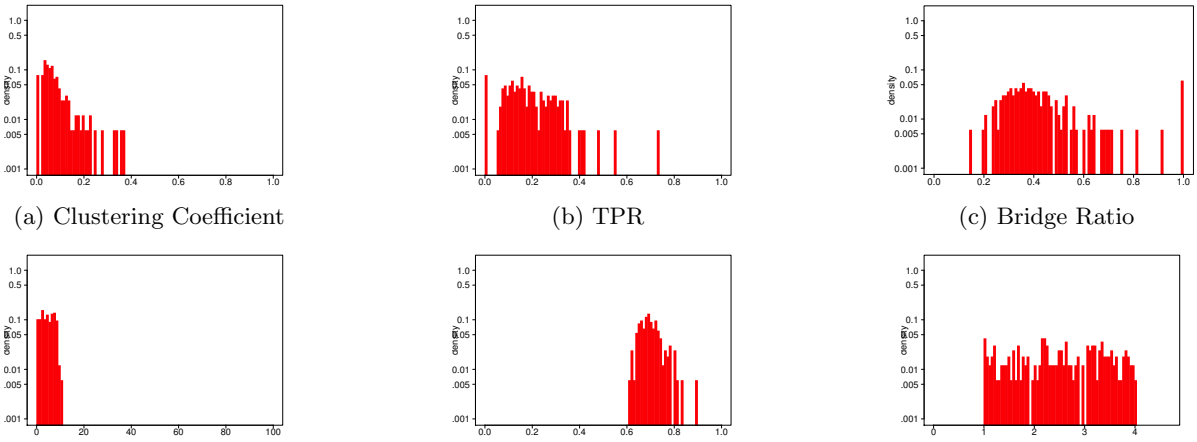
Figure 11: Distribution of the statistical indicators for the LFR2 graph.



(a) Clustering Coefficient (b) TPR (c) Bridge Ratio

(d) Diameter (e) Conductance (f) log10(Size)

Figure 12: Distribution of the statistical indicators for the LFR4 graph.



(a) Clustering Coefficient (b) TPR (c) Bridge Ratio

(d) Diameter (e) Conductance (f) log10(Size)

Figure 13: Distribution of the statistical indicators for the LFR5 graph.