Comparing overlapping properties of real bipartite networks

Fabien Tarissan[1]

Sorbonne Universités, UPMC Université Paris 6 and CNRS, UMR 7606, LIP6, Paris

Abstract. Many real-world networks lend themselves to the use of graphs for analysing and modelling their structure. But such a simple representation has proven to miss some important and non trivial properties hidden in the bipartite structure of the networks. Recent papers have shown that overlapping properties seem to be present in bipartite networks and that it could explain better the properties observed in simple graphs. This work intends to investigate this question by studying two proposed metrics to account for overlapping structures in bipartite networks. The study, conducted on four dataset stemming from very different contexts (computer science, juridical science and social science), shows that the most popular metrics, the clustering coefficient, turns out to be less relevant that the recent redundancy coefficient to analyse intricate overlapping properties of real networks.

Keywords: Complex networks, Bipartite graphs, Social networks, Overlapping

1 Introduction

Many complex networks lend themselves to the use of graphs for analysing and modelling their structure. Usually, vertices of the graph stand for the nodes of the network and the edges between vertices stand for (possible) interactions between nodes of the network. This approach have proven to be useful to identify non trivial properties of the structure of networks in very different contexts, ranging from computer science (the Internet, peer-to-peer networks, the web), to biology (protein-protein interaction networks, gene regulation networks), social science (friendship networks, collaboration networks), linguistics, economy, etc. [17, 4, 12, 2, 6, 13, 1, 14].

Although useful, such a simple representation is not particularly close to the real structure of most of real networks. If one considers for instance actor networks which link actors performing in the same movies [17, 11] or co-authoring networks which link authors publishing together [11, 12], one would rather relate actors to the movies they performed in and authors to their papers. This observation led the community to use *bipartite graphs* instead, i.e. graphs in which nodes can be divided into two disjoint sets, $\top$ (e.g. movies) and $\bot$ (e.g. actors), such that every link connects a node in $\top$ to one in $\bot$. Bipartite graphs are

a fundamental object which has proven to be very efficient for both the analysis [5, 16, 1, 14] and the modelling [3, 15] of complex networks as it is able to reveal patterns that could not have been detected on simple graphs.

In a recent work [15], this framework has been investigated in an attempt to propose, for the first time, a bipartite model of the Internet topology. It relies on recent developments in topology discovery [8, 7] that allows for revealing two layers in the Internet structure. The model remains simple: it only takes as input the node degree sequence for both layers and randomly generates a bipartite graph respecting those distributions. The paper showed that, despite the simplicity of the model, realistic network properties, such as high local density and non trivial correlations among properties of the nodes of the lower layer, emerge naturally. But it also showed that the model fails in reproducing the overlapping observed in the two-layer structure.

The present paper extends the analysis of overlapping structures to a wide variety of networks and tries to identify how bipartite metrics can account for those complex properties. In particular, it investigates whether two recently proposed metrics, namely the *bipartite clustering coefficient* and *bipartite redundancy coefficient*, are relevant for explaining the observed overlaps.

The remaining of the paper is organised as follow: Section 2 will review the technical background necessary for going throughout the paper; Section 3 will present the main obtained results and finally Section 4 will conclude the paper and open on new perspectives.

2 Background

In this section, we introduce the required background for the remainder of the paper. First, we focus on the different dataset (Section 2.1) we used in this study. Then, we recall the necessary definitions of the bipartite graph framework and its related metrics (Section 2.2).

2.1 Dataset

As stated in the introduction, many real networks exhibit a complex structure that involves several layers. In order to be as general as possible in the present study, we used a wide variety of networks presenting a two-level structure. We chose to focus on an infrastructure network (Internet), a juridical network (International Criminal Court decisions network) and two social networks (a co-publication network and a network composed of YouTube users). Here below we describe the four dataset and precise, for each one, the meaning of the upper layer ($\top$ nodes) and the lower layer ($\bot$ nodes):

Internet [15, 8]: In this network, $\bot$ nodes stands for Internet routers and $\top$ nodes indicates the presence of Ethernet switches whose purpose is to induce indirect connections among routers. The dataset used in this study corresponds to a measurement campaign conducted in September 2006.

ICC [14]: This dataset describes the juridical decisions taken by the International Criminal Court (ICC) in the Lubanga case. Here, $\perp$ nodes stands for the juridical decisions made by the judges and $\top$ nodes for the articles of the Rome Statute they invoked. The dataset was extracted from a public server in March of 2013.

Publications [10]: This network describes the scientific collaboration among researchers through co-published papers. It is based on preprints posted to the *Condensed Matter* section of ARXIV E-Print Archive between 1995 and 1999. In this network, $\perp$ nodes stands for authors and $\top$ nodes for articles.

YouTube [9]: This dataset describes some characteristics of YOUTUBE users. It has been collected in 2007 and show the relation between users ($\perp$) and their membership ($\top$).

As we will see further (see Section 3.1 in particular), although the nature of those networks are very different, their two-level structure share some particular and non trivial properties, among which the classical heterogeneous distribution of the degree of the nodes. Note that the size of the networks varies from thousands of nodes to hundred of thousand of nodes. For this reason, all the distributions that we will study further will be normalised by the size of the networks in order to ease the comparisons.

2.2 Bipartite graphs

Bipartite graphs – also referred to sometime by two-mode networks – are triplets $G_b = (\top, \perp, E_b)$, where $\top$ is the set of *top* nodes (the papers in the PUBLICATION dataset for instance), $\perp$ the set of *bottom* nodes (the authors), and $E_b \subseteq \top \times \perp$ the set of links between $\top$ and $\perp$ (that relate the papers to their authors in our example). We denote by $n_\top$ (resp. $n_\perp$) the number of top nodes (resp. bottom nodes) and by m_{bip} the number of links.

Compared to standard graphs, nodes in a bipartite graph are separated in two disjoint sets, and the links are always between a node in one set and a node in the other set. Note that from a given bipartite graph, one can always induce a corresponding simple graph by a $\perp$-projection. In the case of the PUBLICATION network, it would generate a simple graph in which nodes are authors and a link relates two authors if they have published a joint paper. This would allow to reuse all the metrics defined for standard graphs.

But we can also compute specific metrics for bipartite graphs, such as $k_\top$ (resp. $k_\perp$) the average degree of top nodes (resp. bottom nodes), $d_\top^+$ (resp. $d_\perp^+$) the maximal degree observed in top nodes (resp. bottom nodes) and $\delta_\mathfrak{b} = \frac{m_\mathfrak{b}}{n_\top . n_\perp}$ the density of the bipartite graph.

Those are natural extensions of standard metrics defined for simple graphs. But for more intricate properties, it can be tedious to propose a "natural" definition. This is the case for the local density in the graph (more or less the density around a node) which is usually captured by the clustering coefficient.

The reason for the difficulty in defining such an extension is that it relies on the presence of triangles which does not exist in bipartite graphs. As suggested in [5], one can however rely on the following coefficient that tends to capture the overlapping between the neighbourhood of two nodes of $\top$. Let $N_\top(u)$ for $u \in \top$ denote the set of neighbours (i.e. bottom nodes u is linked to) and $N_\bot(u)$ the dual definition for $\bot$ nodes. Then we define:

$$\texttt{cc}_\top(u,v) = \frac{|N_\top(u) \cap N_\top(v)|}{|N_\top(u) \cup N_\top(v)|}. \tag{1}$$

This coefficient is interesting as it captures the relative overlap between neighbourhoods of top nodes, i.e. $\texttt{cc}_\top(u,v)$ is equal to 1 if the neighbourhood of u and v intersects exactly, to 0 if they do not share any neighbour. From this coefficient, it becomes natural to define the clustering coefficient related to a specific $\top$ node v. This is given by:

$$\texttt{cc}_\top(v) = \frac{\sum_{u \in N_\bot N_\top(v)} \texttt{cc}_\top(u,v)}{|N_\bot N_\top(v)|}. \tag{2}$$

This coefficient enables in particular to study the distribution of this property over the top nodes as well as its correlation with the degree or other properties. Then one can naturally compute the *bipartite top clustering coefficient* $\texttt{cc}_{\texttt{bip}}$ of G_b as the average value of $\texttt{cc}_\top(v)$ over all the nodes v of $\top$. More formally:

$$\texttt{cc}_{\texttt{bip}}(G_b) = \frac{1}{|\top|} \sum_{v \in \top} \texttt{cc}_\top(v). \tag{3}$$

However it has been shown in [5] that this coefficient might miss some important properties of the overlapping between $\top$ nodes in the bipartite structures. This is why the authors suggested to use the *redundancy coefficient* $\texttt{rd}_\top(v)$ of a node v which focuses on the impact of removing v as regard the $\bot$-projection. Intuitively, a high value of the coefficient indicates that two $\bot$ nodes v relates are likely to be related by another $\top$ node. Formally, the coefficient is given by:

$$\texttt{rd}_\top(v) = \frac{|\{\{u,w\} \in N_\top(v)^2 \text{ s.t. } \exists v' \neq v, (v',u) \in E_b \text{ and } (v',w) \in E_b\}|}{\frac{|N_\top(v)|(|N_\top(v)|-1)}{2}}. \tag{4}$$

Following this definition, we can derive naturally the redundancy coefficient $\texttt{rd}_{\texttt{bip}}$ of the bipartite graph G_b, defined as the average value of the former coefficient over all $\top$ nodes. More formally:

$$\texttt{rd}_{\texttt{bip}}(G_b) = \frac{1}{|\top|} \sum_{v \in \top} \texttt{rd}_\top(v). \tag{5}$$

	Internet	ICC	Publication	YouTube
$n_\top$	10 224	713	22 015	30 087
$n_\bot$	9 758	1 360	16 726	94 238
$m_\mathtt{b}$	25 422	6 670	58 595	293 360
$\delta_\mathtt{b}(*10^{-3})$	0.26	6.88	0.16	0.10
$k_\top$	2.5	9.4	2.7	9.8
$k_\bot$	2.6	4.9	3.5	3.1
$d^+_\top$	58	250	18	7 591
$d^+_\bot$	41	81	116	1 035

Table 1. Global properties of the bipartite structure of the dataset

	Internet	ICC	Publication	YouTube
$\mathtt{cc}_\mathtt{bip}$	0.32	0.15	0.39	0.16
$\mathtt{rd}_\mathtt{bip}$	0.11	0.69	0.63	0.33

Table 2. Value of the overlapping coefficients of the bipartite structures

3 Analysis of the bipartite structure

The purpose of this section is to analyse the overlapping observed in the bipartite structure of the four dataset presented in Section 2.1. We will focus in particular to the two metrics that have been proposed to account for such a topological property, namely the bipartite clustering coefficients and the bipartite redundancy coefficients (referred to further simply as clustering and redundancy). First, we start by looking at some global and standard statistics defined for bipartite graphs (Section 3.1). Then we turn to the overlapping properties and study distributions and correlations among the different metrics (Section 3.2).

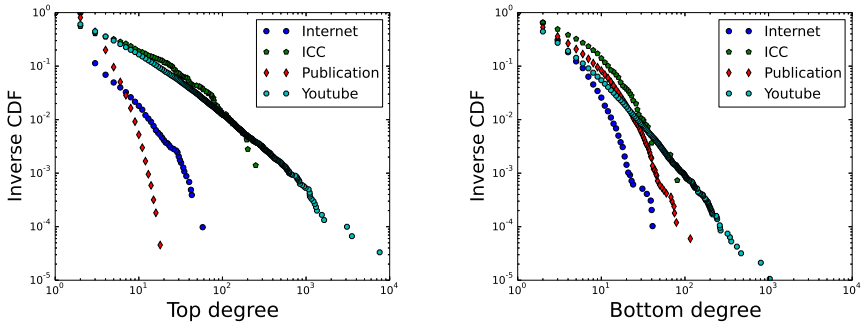
3.1 A global perspective

The first statistics we focus on concern some basic properties observed in most real-world networks, formally presented in the previous section. Table 1 presents the results for the four dataset of Section 2.1. As expected, all usual observations made on real-world networks stand also for the networks under study. In particular the graph is sparse (on the order of magnitude of 10^{-4}) and the maximal degree is several orders of magnitude higher that the average degree, which indicates usually some heterogeneity in the degree of the nodes.

This is confirmed by the inverse cumulative distribution of the degree of the nodes (both $\top$ and $\bot$) presented in Figure 1. It clearly shows a heavy-tail distribution for all the four dataset and the nodes of the two layers.

3.2 Analysis of the overlapping structure

We focus now more precisely to the core of the analysis related to the overlapping in the bipartite structure. First, Table 2 presents the global values of the



(a) Top nodes

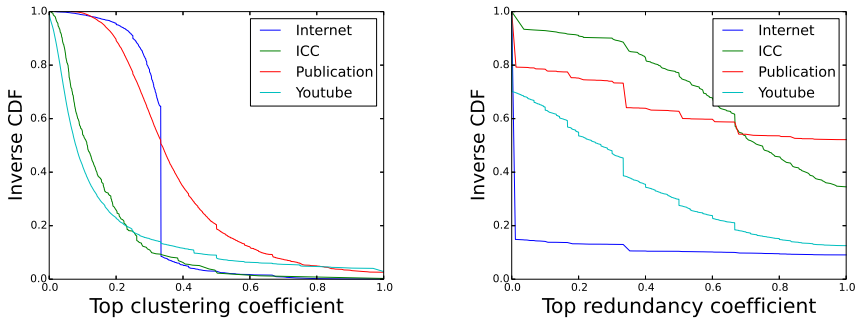
(b) Bottom nodes

Fig. 1. Inverse CDF of the node degree distribution

two coefficients computed for all the dataset. It shows that, although the two coefficients intend to capture the same property (overlapping patterns), they strongly differ on each dataset. The most obvious case is the ICC since the clustering coefficient is quite low (0.15) but the redundancy is very high (0.69). For the other dataset the gap is less important but we can notice that the higher coefficient depends on the dataset thus showing that no general behaviour can be drown here.

Those global average metrics do not allow for a detailed comprehension of the coefficients. Fortunately, we can compute them for each of the $\top$ nodes in the networks. This allows to study several properties related to it such as the distribution of the coefficients. Figure 2 presents the inverse cumulative distribution of the clustering (Figure 2(a)) and the redundancy (Figure 2(a)). We can observe that the distributions of the two coefficients are very different. For the clustering, the plot shows that the decrease of the value is very sharp and for low values. The majority of $\top$ nodes have indeed a small clustering coefficient. This indicates that the overlapping, according to this metrics, is not particularly important in the networks.

For the redundancy, the behaviour is different. Except for the Internet case, for which one can observe a sharp decrease, the value is uniformly distributed among the nodes. As opposed to the clustering, this seems to indicate on the contrary that some overlaps are present in three over four dataset. Note for instance that the fraction of $\top$ nodes having a redundancy of 1 is non negligible: 9% in Internet, 13% in the YouTube case, 46% in the ICC network and 52% in the Publication network. Taking this last case as an example it means that, for more than half of the articles of the Publication network, every authors have also published together at least one other article. This indicates a strong overlapping in the network which, in the case of co-publication networks, is not surprising but is not captured by the clustering coefficient.



(a) Clustering (b) Redundancy

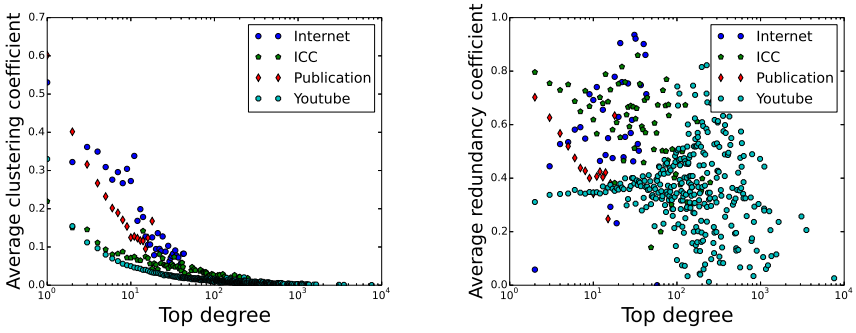
Fig. 2. Inverse CDF of the clustering and redundancy coefficients

The distribution shown above is interesting but it does not help to understand why some coefficients are high and other low. In order to understand better the situation, we show Figure 3 the correlation between the degree of a $\top$ node and the value of its clustering (Figure 3(a)) or its redundancy (Figure 3(b)). More precisely, a (x, y) dot in the plots means that the average value of the coefficient for nodes having degree x is y. Figure 3(a) shows a very interesting fact: the value of the clustering seems to be completely governed by the degree of the corresponding node. The higher the degree, the lower the clustering. Such a correlation makes the interest of the coefficient weak since it seems derivable from the degree of the nodes. On the contrary, Figure 3(b) does not present such a correlation, except for the PUBLICATION network for which one observe a similar behaviour. The notion engulfed in the redundancy coefficient seems then, to that regard, contain more information than simpler local properties.

4 Conclusion

In this paper, we studied the overlapping properties observed in the bipartite structure of different networks exhibiting a two-level structure. The main concern of the study was to discriminate between two recently proposed metrics to account for such properties, namely the clustering coefficient and the redundancy coefficient.

By analysing the structure of 4 networks stemming from very different contexts, we showed that the notion captured by the clustering coefficient turns out to be quite poor as it is closely related to the simple degree of the node. On the contrary, the behaviour of the redundancy coefficient is totally unpredictable regarding local properties such as the degree. The value of the coefficient is not related to simple local properties of the nodes, at least in 3 of the 4 dataset of the study.



(a) Clustering (b) Redundancy

Fig. 3. Correlation between the degree and the overlapping coefficients

Understanding the characteristics of the bipartite structure of real networks are fundamental for several reasons. First, as shown in several studies, such structures help understanding non trivial properties of simple networks (see [15] for instance). But more importantly it has been shown to be a better support for models, enabling in particular to generate random structures closer to real ones than most of classical models [15].

To that regard, the present work opens the way to several improvements in recently proposed models. It shows in particular that one could improve bipartite models by integrating such a property in the model, which has not been done so far. One way to achieve this goal would be to *encode* the redundancy in an artificial third level and control the coefficient by randomly permuting links in such a tripartite structure. We let such an investigation as a further work.

Acknowledgement

This work is partly funded by the National Center for Scientific Research (CNRS) through the PEPS Project "DoRé".

References

1. Yong-Yeol Ahn, Sebastian E Ahnert, James P Bagrow, and Albert-László Barabási. Flavor network and the principles of food pairing. *Scientific reports*, 1, 2011.
2. Stefano Battiston and Michele Catanzaro. Statistical properties of corporate board and director networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):345–352, 2004.
3. Jean-Loup Guillaume and Matthieu Latapy. Bipartite graphs as models of complex networks. *Physica A: Statistical Mechanics and its Applications*, 371(2):795–813, 2006.

4. Ramon Ferrer i Cancho and Richard V Solé. The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1482):2261–2265, 2001.
5. Matthieu Latapy, Clémence Magnien, and Nathalie Del Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31–48, 2008.
6. Fabrice Le Fessant, Sidath Handurukande, A-M Kermarrec, and Laurent Massoulié. Clustering in peer-to-peer file sharing workloads. In *Peer-to-Peer Systems III*, pages 217–226. Springer, 2005.
7. P. Mérindol, B. Donnet, O. Bonaventure, and J.-J. Pansiot. On the impact of layer-2 on node degree distribution. In *Proc. ACM/USENIX Internet Measurement Conference (IMC)*, November 2010.
8. P. Mérindol, V. Van den Schriek, B. Donnet, O. Bonaventure, and J.-J. Pansiot. Quantifying ASes multiconnectivity using multicast information. In *Proc. ACM/USENIX Internet Measurement Conference (IMC)*, November 2009.
9. Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07)*, San Diego, CA, October 2007.
10. M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):404–409, 2001.
11. Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions. *Physics Reviews E*, 64, 2001.
12. Mark EJ Newman, Duncan J Watts, and Steven H Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(Suppl 1):2566–2572, 2002.
13. Christophe Prieur, Dominique Cardon, Jean-Samuel Beuscart, Nicolas Pissard, and Pascal Pons. The stength of weak cooperation: A case study on flickr. *arXiv preprint arXiv:0802.2317*, 2008.
14. Fabien Tarissan and Raphaëlle Nollez-Goldbach. The network of the international criminal court decisions as a complex system. In A. Sanayei, I. Zelinka, and O. E. Rossler, editors, *ISCS 2013: Interdisciplinary Symposium on Complex Systems*, volume 8 of *Emergence, Complexity and Computation*, pages 225–264. Springer, 2013.
15. Fabien Tarissan, Bruno Quoitin, Pascal Mérindol, Benoit Donnet, Jean-Jacques Pansiot, and Matthieu Latapy. Towards a bipartite graph modeling of the internet topology. *Computer Networks*, 57(11):2331–2347, 2013.
16. Michele Tumminello, Salvatore Miccichè, Fabrizio Lillo, Jyrki Piilo, and Rosario N Mantegna. Statistically validated networks in bipartite complex systems. *PloS one*, 6(3):e17994, 2011.
17. Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *nature*, 393(6684):440–442, 1998.