Revealing intricate properties of communities in the bipartite structure of online social networks

Raphaël Tackx, Jean-Loup Guillaume, Fabien Tarissan

Revealing intricate properties of communities in the bipartite structure of online social networks

Raphal Tackx*, Jean-loup Guillaume†, Fabien Tarissan*
* Sorbonne Université, UPMC Univ Paris 06, UMR 7606, LIP6,F-75005, Paris, France
CNRS, UMR 7606, LIP6, F-75005, Paris, France
† L3I, Université de La Rochelle, Av. Michel Crepeau, 17042 La Rochelle, France

***Abstract*—Many real-world networks based on human activities exhibit a bipartite structure. Although bipartite graphs seem appropriate to analyse and model their properties, it has been shown that standard metrics fail to reproduce intricate patterns observed in real networks. In particular, the overlapping of the neighbourhood of communities is difficult to capture precisely. In this work, we tackle this issue by analysing the structure of 4 real-world networks coming from online social activities. We first analyse their structure using standard metrics. Surprisingly, the clustering coefficient turns out to be less relevant than the redundancy coefficient to account for overlapping patterns. We then propose new metrics, namely the *dispersion* and the *monopoly* coefficients, and show that they help refining the study of bipartite overlaps. Finally, we compare the results obtained on real networks with the ones obtained on random bipartite models. This shows that the patterns captured by the redundancy and the dispersion coefficients are strongly related to the real nature of the observed overlaps.**

Keywords*—*online social networks, complex networks, bipartite graphs, overlapping, communities

I. Introduction

Many real-world networks – also referred to as *complex networks* – lend themselves to the use of graphs for analysing and modelling their structure. Usually, the 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. We can cite for instance computer networks (like the Internet, peer-to-peer systems, the web) [1], [2], biological networks (protein-protein interaction networks, metabolic processes) [3], [4], social networks (friendship networks, co-publication networks) [3], [5], legal networks [6], linguistics [7], economy [8], etc.

However, even though this representation based on graphs is relevant, it is often too simplistic to account for the inherent complexity of most networks. Indeed, if we consider for instance an actor-movie network [3], [9] which relates actors to movies, or a co-publication network [9], [5] which connects authors to publications, it is then natural to distinguish formally the two kind of vertices as two disjoint sets: actors on one side, movies on the other and authors on one side and publications on the other. This observation has led the scientific community to use *bipartite graphs* to represent such a constrained structure. Indeed in bipartite, also referred as *2-mode networks*, the set of nodes is composed of two disjoint parts, $\top$ and $\bot$ (for example movies and actors), such that edges only relate vertices of different sets. It has been shown that this fundamental object is reliable both for analysing [10], [11], [12] and modeling [13], [14] the structure of real-world networks.

In a recent paper [14], this approach has been exploited in order to propose for the first time a bipartite model of the Internet topology. The model is general enough to be relevant for any network presenting a bipartite structure as it only takes as parameters the degree sequence of the nodes in the two disjoint sets. The paper showed that despite the simplicity of the model, non trivial realistic properties of the Internet topology emerge naturally. But it also showed that the model fails in reproducing the overlapping observed in the bipartite structure. This overlapping arises when two $\bot$ nodes are related to several $\top$ nodes (two authors publishing more than one paper together for instance). The observation drawn from the study conducted on Internet topology has been extended to a wide variety of different networks in [15], showing that overlaps are common in complex bipartite networks. Understanding the structure of the overlaps in bipartite networks is thus a key concern in the processes of improving our ability to analyse and model such networks.

The purpose of the present paper is precisely to address the question of the real nature of overlapping patterns observed in real networks. To do so, we analyse the structure of 4 different real-world networks coming from online social activities. This choice is motivated by the fact that such networks exhibit a genuine bipartite structure since online platforms usually propose to the users the ability to gather into communities. To that regard, these networks are naturally represented by bipartite graphs with $\bot$ nodes defining the users and $\top$ nodes depicting the communities they belong to. We then analyse the overlapping patterns through the use of both standard metrics and new metrics proposed in this paper.

The remaining of the paper is organised as follow. First, we detail the different datasets on which we conducted the study and give proper definition of the different metrics we used to analyse their structure (Section II). Then we present the results obtained and address in particular the variety of overlaps in real networks as well as in random models (Section III). We finally conclude the paper by giving some perspective on future work (Section IV).

II. Datasets and definitions

In this section, we first present the different datasets we used in this study (Section II-A) before describing the existing

(Section II-B) and new (Section II-C) metrics we used to analyse them.

A. Datasets

As explained before, many real-world social network exhibit an intricate structure involving two layers. In order to have an approach that allows to draw general conclusions, we relied on a wide variety of social networks involving communities on a broad sense. We focused in particular on two membership networks (FLICKR, LIVEJOURNAL) and two publications networks (CITEULIKE, WIKIPEDIA). Here below, we briefly describe them explaining what $\top$ nodes and $\bot$ nodes stand for in these contexts:

- LIVEJOURNAL [16]: This dataset concerns a website where users are freely allowed to post information using a blogging system. It also allows people to declare friendships to each other, and to create and join groups. Here, $\top$ nodes stand for groups and $\bot$ nodes for users. It involves approximately 1 million users and more than 600,000 groups.
- CITEULIKE [17]: This dataset comes from a large online library which allows users to share, cite or tag scientific publications. Here we rely on about 150,000 tags and 700,000 publications, considering that $\top$ nodes are the tags and $\bot$ nodes are the publications.
- WIKIPEDIA [18]: Here the $\bot$ nodes are WIKIPEDIA articles and $\top$ nodes are categories that group pages on similar topics. For computational reason, we did not use the whole network and rather focused on a large connected component [1]. This sub-network involves 3 millions articles and almost 500,000 subcategories.
- FLICKR [16]: This dataset is composed of 100,000 groups and 400,000 users (respectively $\top$ and $\bot$ nodes) from the FLICKR website that allows hosting and sharing pictures.

Detailed statistics on those datasets are provided extensively in Section III.

B. Metrics for bipartite graphs

In order to best account for the real structure of the datasets presented above, we rely in this study on *bipartite graphs*. Bipartite graphs are triplets $G_b = (\top, \bot, E_b)$, where $\top$ is the set of *top* nodes (e.g., the groups in LIVEJOURNAL), $\bot$ the set of *bottom* nodes (e.g., the users in LIVEJOURNAL), and $E_b \subseteq \top \times \bot$ the set of links between $\top$ and $\bot$ (that relate the users to their groups). We denote by $n_\top$ (resp. $n_\bot$) 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 ($E_b \subseteq \top \times \bot$), see Fig. 1. Note that for some datasets the information is richer since we also have links between $\bot$ nodes (resp. $\top$ nodes), for instance links between articles (resp. a hierarchy of categories) in WIKIPEDIA. These additional links are very rich and should be used. However since we focus on the overlapping patterns we decided to discard this extra information.

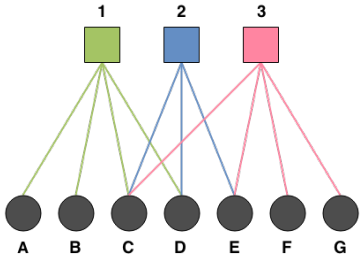


Fig. 1. Example of a bipartite graph.

There are many natural extensions of standard metrics defined for simple graphs to the case of bipartite graphs, such as $d_\top(u)$ (resp. $d_\bot(v)$) the degree (i.e. the number of neighbors) of a top node u (resp. bottom node v), $k_\top$ (resp. $k_\bot$) the average degree of top nodes (resp. bottom nodes), $d^+_\top$ (resp. $d^+_\bot$) the maximal degree of top nodes (resp. bottom nodes) and $\delta_\mathtt{b} = \frac{m_{\mathtt{bip}}}{n_\top . n_\bot}$ the density of the bipartite graph.

But for more intricate properties, it can be tedious to propose a "natural" definition. This is for instance the case of the local density in the graph, which is defined as the density of the induced subgraph by the neighbours of a node in simple graphs. In bipartite graphs, this local density usually tries to capture how the neighbourhoods of $\top$ nodes tend to overlap each other. Some definitions have been already proposed to study this property in bipartite graphs [19], [10], [20], [21], [22]. In the present work, we rely on two of those metrics to cope with the local density : the *bipartite clustering coefficient* and the *redundancy coefficient* [10]. The first one focuses on the intersection of the neighbourhood of two $\top$ nodes, while the second one focuses on the impact of removing v as regard to the relations between $\bot$ nodes.

More formally, let $N_\bot(u)$ denote the set of neighbours of a node $u \in \top$ (i.e., bottom nodes u is linked to) and $N_\top(v)$ the dual definition for a $\bot$ node v. We can then define the bipartite clustering coefficient between two top nodes by:

$$\mathtt{cc}(u,v) = \frac{|N_\bot(u) \cap N_\bot(v)|}{|N_\bot(u) \cup N_\bot(v)|}. \quad (1)$$

This coefficient is equal to 1 if the neighbourhoods of u and v are equal (complete overlap), 0 if they have no $\bot$ nodes in common. Then, it is natural to derive the clustering coefficient of u $\mathtt{cc}(u)$ as the average value of $\mathtt{cc}(u,v)$ for all $v \in \top$ that share at least one common neighbour with u. Finally we can derive $\mathtt{cc}_{\mathtt{bip}}$, the clustering coefficient of the graph G_b, as the average value for all $\top$ nodes [2]

As regard the *redundancy coefficient* of a $\top$ node u, it is formally defined as:

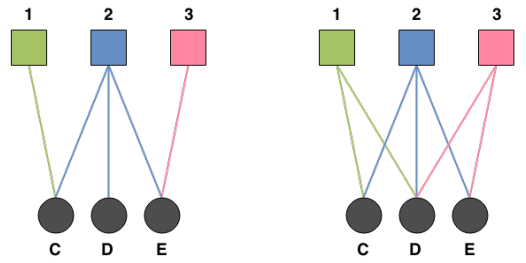
[1] We chose to focus on the connected component containing the *Network Protocol* category

[2] Note that for this coefficient, as well as for the next ones, the dual notion for bottom nodes can be derived, although we will not use it in the present article.

$$\texttt{rd}(u) = \frac{|\{\{v,w\} \in N_\perp(u)^2 \text{ s.t. } N_\top(v) \cap N_\top(w) \neq \{u\}\}|}{\binom{|N_\perp(u)|}{2}}. \quad (2)$$

Intuitively a high value of the coefficient indicates that two $\perp$ nodes that u relates are likely to be also related by another $\top$ node (two users belonging to a group u also share another group in common), thus revealing some overlapping pattern in the structure. As for the clustering coefficient, we can derive the redundancy coefficient $\texttt{rd}_{\texttt{bip}}$ of the bipartite graph G_b, as the average value of the former coefficient over all $\top$ nodes.

Looking at the examples provided in Fig. 2, we can assess the relevance of the redundancy coefficient for identifying overlapping pattern. In Fig. 2a in which there is no overlap for node 2, one can check that $\texttt{rd}(2) = 0$ while in Fig. 2b, the value of the coefficient is $\frac{2}{3}$ which reflects the fact that 2 over the 3 possible relations between nodes C, D and E are not affected by the removal of node 2 due to overlapping with nodes 1 and 3.



(a) Structure with no overlap

(b) Structure with overlap

Fig. 2. Two bipartite graphs depicting a different structure around node 2.

C. Two new metrics to describe the overlap

In order to refine those classical metrics, we now propose two new coefficients, related to the overlapping structure in bipartite graphs. We first define the *dispersion coefficient* which focus on how are distributed the links among the neighbours of $u \in \top$. More precisely, we compute the maximal number of different communities that could be related to the set of neighbours of u. Let $N_\top(u)$ denote the set of $\top$ neighbours of a node $u \in \top$ (i.e., the set of all $\top$ nodes that share at least one bottom node with u). The dispersion coefficient $\texttt{disp}(u)$ is then defined by:

$$\texttt{disp}(u) = \frac{|N_\top(u)| - 1}{\sum_{v \in N_\perp(u)} (|N_\top(v)| - 1)} \quad (3)$$

Intuitively, the dispersion measures whether members a group tend to share same other groups or not. If the dispersion of a $\top$ node u is complete, then the size of the set of all communities related to its $\perp$ neighbours ($|N_\top(u)| - 1$) will match exactly the sum of their degree (minus 1). Thus the value of the coefficient is equal to 1. This is the case for node 2 in Fig. 2a. This value is directly correlated to the fact that the redundancy of this node is equal to 0, no overlap implying complete dispersion of the links. On the other hand, when some overlapping pattern are present (see Fig. 2b), the dispersion tends to decrease: $\texttt{disp}(2) = \frac{|\{1,3\}|}{1+2+1} = \frac{1}{2}$.

Finally, we propose another metric which accounts for the specificity of degree-1 $\perp$ nodes that are particularly numerous in real-world networks, thus impacting the analysis of overlapping patterns. We therefore define the *monopoly coefficient* as the proportion of degree-1 neighbours of a $\top$ node u. Formally:

$$\texttt{mon}(u) = \frac{|\{v \in N_\perp(u) \text{ s.t. } d_\perp(v) = 1\}|}{|N_\perp(u)|} \quad (4)$$

As for the former coefficient, it is naturally possible to derive the dispersion, or the monopoly, of a bipartite graph as the average value of the corresponding coefficients over all $\top$ nodes. As we will see further in the next section, these metrics are able to refine the analysis provided by the standard metrics and help understanding the true nature of the observed overlaps in bipartite networks.

III. Analysis of the bipartite structure

The purpose of the present section is to understand the behaviour of the different metrics computed on the four datasets presented previously. We start by providing general statistics for the different networks (Section III-A). Then we detail the analysis by investigating in particular the relevance of the traditional clustering and redundancy coefficients to characterise overlapping patterns in bipartite networks (Section III-B), whether the new proposed metrics are able to refine the analysis (Section III-C) and how random models affect the structure of the networks as regards to these properties (Section III-D).

As for the former coefficient, it is naturally possible to derive the dispersion (resp. monopoly) of a bipartite graph as the average value of the dispersion (resp. monopoly) coefficients over all nodes, which we denote by $\texttt{disp}_{\texttt{bip}}$ (resp. $\texttt{mon}_{\texttt{bip}}$).

A. Global properties

	CiteULike	LiveJournal	WikiPedia	Flickr
$n_\top$	153,3 K	664,4 K	484,5 K	103,6 K
$n_\perp$	731,8 K	1,15 M	3,13 M	396 K
$\delta_{\texttt{b}}(*10^{-5})$	2.1	0.9	0.7	20.8
$k_\top$	15.02	10.79	22.31	82.46
$k_\perp$	3.20	6.24	3.46	21.58
$d^+_\top$	153 K	149 K	36 K	35 K
$d^+_\perp$	1,3 K	682	138	2.2 K
$\texttt{cc}_{\texttt{bip}}$	0.138	0.117	0.063	0.055
$\texttt{rd}_{\texttt{bip}}$	0.521	0.703	0.387	0.646
$\texttt{disp}_{\texttt{bip}}$	0.725	0.842	0.705	0.769
$\texttt{mon}_{\texttt{bip}}$	0.071	0.070	0.088	0.148

TABLE I. Global properties of the bipartite structure of the datasets.

We first focus on some global statistics of bipartite graphs as presented in the previous section. Table I presents the results for the four datasets. It shows that the networks under investigations present standard properties commonly observed in real-world networks [13]. In particular one can notice that the networks are globally sparse ($\delta_{\texttt{b}}$) while local densities are order of magnitudes higher ($\texttt{cc}_{\texttt{bip}}$ and $\texttt{rd}_{\texttt{bip}}$). Besides, the order of magnitude between the average degrees ($k_\top$ and $k_\perp$) and the maximal degrees ($d^+_\top$ and $d^+_\perp$) indicates some

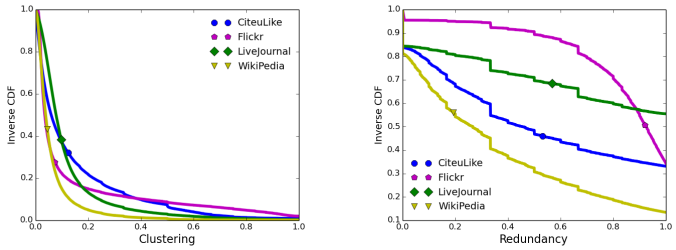
heterogeneity in the degree distribution of both $\top$ and $\bot$ nodes [3].

As regard the two classical metrics usually used to capture overlaps in bipartite structures ($\texttt{cc}_{\texttt{bip}}$ and $\texttt{rd}_{\texttt{bip}}$), one can notice that they strongly differ. By relying on the clustering coefficient, those networks seems to present few overlaps while, on the contrary, the redundancy coefficient tends to indicates the presence of overlaps. The FLICKR case is to that regard eloquent. This raises the question of understanding which one of the two coefficients is the more relevant to depict the presence of overlaps in the structure. This question will be investigated more deeply in Section III-B.

If one focus now on the new metrics ($\texttt{disp}_{\texttt{bip}}$ and $\texttt{mon}_{\texttt{bip}}$), one can notice that the average monopoly is very low for each network, indicating that degree-1 $\bot$ nodes seem well distributed in the communities. On the contrary, the average dispersion is high. This seems to contradict the redundancy coefficient as it tends to indicate a lack of overlaps in the structure. But such an aggregated value is difficult to analyse and a more detailed discussion is provided in Section III-C.

B. Standard bipartite metrics

The global statistics presented in the previous section do not allow to grasp the diversity of the possible situations for all $\top$ nodes. In order to refine this first analysis, we then compute the value of the different coefficients for all $\top$ nodes, which allows to study the distribution of the values as well as some correlations between the metrics.



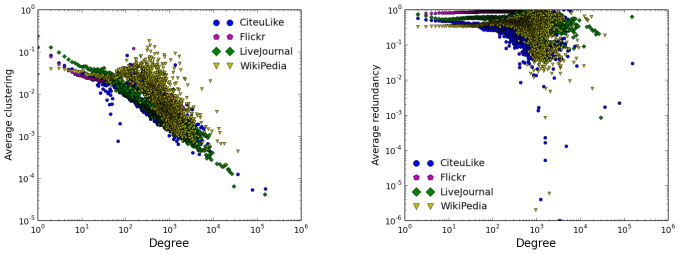
(a) Clustering coefficient (b) Redundancy coefficient

Fig. 3. Inverse cumulative distribution of the clustering and redundancy coefficient.

Fig. 3 presents the inverse cumulative distribution function of the clustering (Fig. 3a) and redundancy (Fig. 3b) coefficients, showing clearly two different distributions. Whereas the distributions sharply decrease for the clustering coefficient, it seems to be quite uniformly distributed for the redundancy coefficient, except for the extreme values (0 and 1) which are highly represented.

More importantly, one can notice that the proportion of poorly clustered nodes is large. More than 75% of the nodes have a value lower than 0.2 in all datasets. The proportion of redundant nodes are in contrast particularly high. The fraction of $\top$ nodes that have a redundancy of 1 is for instance extremely important for all datasets: 17% for WIKIPEDIA, 38% for CITEULIKE and FLICKR, 58% for LIVEJOURNAL. Considering this last case, it means that for more than half of the groups created in LIVEJOURNAL, *every* pair of members belong to (at least) one other group in common. This is surprising, especially since the average numbers of groups a user belong to is very low (6.24, see Table I). This indicates the presence of non trivial overlapping pattern in the networks which is not captured by the clustering coefficient, thus leading to the conclusion that the notion of redundancy is more suited to identify real overlapping patterns in bipartite networks.

[3]Due to space limitation, degree distributions plots are not presented here but they confirm this statement.



(a) Degree vs. Clustering (b) Degree vs. Redundancy

Fig. 4. Correlations between the degree and the clustering/redundancy coefficients.

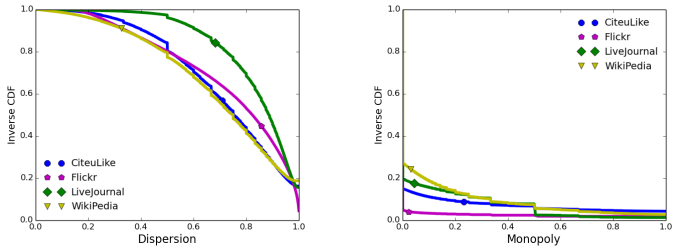
In order to understand why such a difference is observed, we present Fig. 4 the correlation between the degree of a $\top$ node and its clustering or redundancy coefficient. More precisely, we consider for each $\top$ node u a point with coordinate $(\texttt{d}_{\bot}(v), \texttt{disp}(v))$ (Fig. 4a) or $(\texttt{d}_{\bot}(v), \texttt{rd}(v))$ (Fig. 4b). The plot shows the average value for different intervals.

Fig. 4a shows a clarifying fact: the value of the clustering seems to be strongly related to the degree of the node: the higher the degree, the lower the clustering. Such a correlation is not present for the redundancy, Fig. 4b. This strengthens our former conclusion on the relevance of the redundancy to detect overlaps. Indeed, the redundancy seems to be independent of the degree, while the clustering can be derived from it and is thus more related to the *number* of neighbours and less to the *structure* of the relations.

C. Refining the analysis with new metrics

The results presented above focus on the standard metrics proposed as extensions of the notion of *local density* for bipartite graphs. It shows that the notion of redundancy is well suited to detect overlaps in the networks. However, such a notion does not allow to understand how those overlaps are organised around a $\top$ node. The notion of dispersion, as well as the monopoly, proposed in this paper, is an attempt to precise such organisation. This section presents the analysis conducted on the datasets through the use of those two coefficients.

Fig. 5 presents the inverse cumulative distribution of the dispersion (Fig. 5a) and the monopoly (Fig. 5b) coefficients. As suggested by the average values (see Table I), the dispersion is high for large fraction of nodes. In all dataset, more than 80% of the nodes have a value higher than 0.5. This is however coherent with the presence of overlaps around those nodes since 0.5 means than half of the links creates similar relations among $\bot$ and $\top$ nodes, thus inducing overlaps. Yet, these high values contrast with the ones of the redundancy coefficients

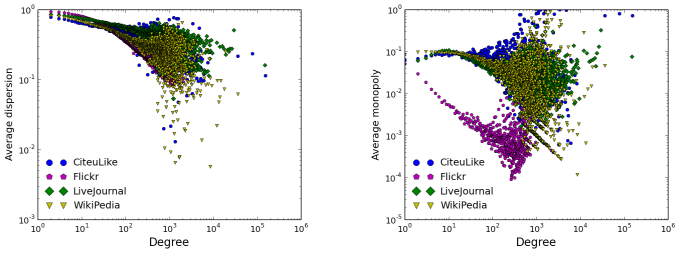


(a) Dispersion coefficient (b) Monopoly coefficient

Fig. 5. Inverse cumulative distribution of the dispersion and monopoly coefficient.

and reduce the importance of the overlaps in the networks. In that sense, it refines the use of redundancy.

As regard monopoly coefficients, their distribution precise our former statement on the fact that it is uniformly distributed. This is obvious in all curves as they all decrease smoothly for the complete range of strictly positive values. However, the reader might notice the high fraction of $\top$ nodes having no monopoly at all. Going back to the global properties of the networks (Table I), it is easily explained by the fact that the number of $\bot$ nodes is way lower than the one of $\top$ nodes. Thus, even if the links are well distributed over the networks, not every top nodes can be related to a degree-1 bottom node, thus leading to a monopoly coefficient of 0.



(a) Degree vs. Dispersion (b) Degree vs. Monopoly

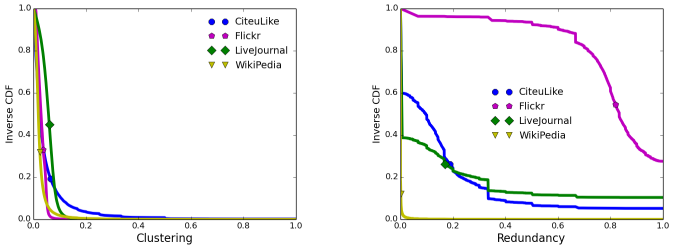
Fig. 6. Correlations between the degree and the dispersion/monopoly coefficients.

As for the clustering and the redundancy, we can precise our analysis by looking at some correlations among the metrics. Fig. 6 presents the correlations of the degree of a $\top$ node with its dispersion and its monopoly. Fig 6a shows a clear tendency: the higher the degree, the lower the dispersion in average. Since low value of dispersion is related to overlaps, this refines our former analysis as we can now precise which kind of top nodes are related for overlaps.

As regard the correlation between the degree and the monopoly, Fig 6b shows that on average the monopoly seems to be independent from the degree, except for FLICKR that displays a clear decrease when the degree increases. One can note that the case of FLICKR is still consistent with our previous statement : since degree-1 nodes cannot be responsible for overlaps in the bipartite structure, it is then coherent with the fact that overlaps are more related to large top nodes.

D. Impact of random model

Having analysed the overlaps in the datasets using two traditional and two new metrics, we now turn to the impact random models have on those coefficients. To do so, we use a variant of the *configuration model* [23] which randomly assigns edges to match a given degree sequence without adding any other expected property. In other word, the generated bipartite graphs have the same number of nodes and links but the links are distributed uniformly at random among $\bot$ and $\top$ nodes, according to their real degree. This section shows the impact such a shuffling have on some of the metrics.

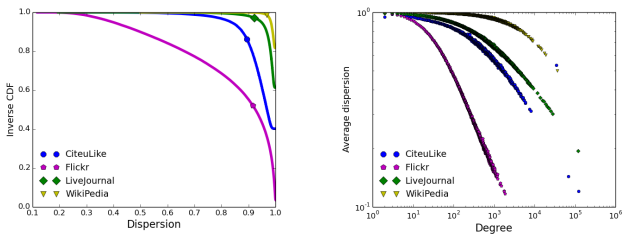


(a) Clustering coefficient (b) Redundancy coefficient

Fig. 7. Inverse cumulative distribution of the clustering and redundancy coefficients for random bipartite graphs.

Fig. 7 shows the inverse cumulative distribution function of the clustering (Fig. 7a) and redundancy (Fig. 7b) coefficients for random networks. As expected, Fig. 7a shows that the randomisation process has little impact on the clustering distribution. Indeed, as shown in Section III-B, the value of the clustering is directly related to its degree. Since the degree sequence is kept unchanged by the model, so does the clustering. This strengthens again our former conclusion to that regard.

Concerning the redundancy, the behaviour is different depending on the dataset on which we apply the model. If the model seems to have completely remove the overlaps in the WIKIPEDIA dataset, it is the complete opposite for the FLICKR network. In this last case, the model seems to have reinforced the presence of the overlaps. If no strict conclusion can be drawn here, the reader might notice that the proportion of high redundancy coefficients is related to the density of the network. The relative density of the FLICKR network (30 times higher that the WIKIPEDIA network) could thus explain the effect of the model.



(a) Dispersion coefficient (b) Degree vs. Dispersion

Fig. 8. Inverse cumulative distribution of the dispersion coefficient and correlation between degree and dispersion for random networks with the same $\top$ and $\bot$ degree distributions.

Finally, we show Fig. 8 the impact of the model to the dispersion coefficient. Fig. 8a presents the inverse cumulative distribution. The plot indicates that the model has increased the values in all cases. This is not surprising since the definition of the dispersion coefficient is related to the notion of expected distribution of the links. In other word, by distributing the links uniformly at random in the networks, the model tends to maximise this coefficient.

Fig. 8b shows the impact the random model has on the correlation between degree and dispersion. Surprisingly, whereas Section II-C clearly identified the degree as directly related to the dispersion in all the dataset, the model shows that this correlation is due to the nature of the network since, when shuffling the real distribution of the links, this correlation is strongly affected. Yet, the tendency identifies previously is still correct: the higher the degree, the lower the dispersion.

IV. Conclusion

In this paper, we studied the overlaps observed in 4 different online social networks exhibiting a bipartite structure. To do so we start by relying on two standard metrics recently proposed to address the notion of local density in bipartite graphs, namely the clustering coefficient and the redundancy coefficient. Our analysis revealed that, the clustering coefficient is surprisingly not particularly able to detect real observed overlaps in networks. To explain this result, we showed that this coefficient is more related to the number of neighbours and less to the structure of the relations.

In order to deepen the study of the overlaps, we proposed two new metrics, namely the dispersion and the monopoly, that complete the analysis. The results show that it help refining the characterisation of the overlaps made by the redundancy coefficient by giving insights on the kind of nodes affected by overlaps. Finally, we applied a random bipartite model in order to assess how the shuffling of the links can affect the properties observed in the present study. It showed in particular that such a randomisation process affect the dispersion of the link, while it keep unchanged the clustering, thus strengthening the interest of this new metric as a valid candidate for the characterisation of the overlaps.

Since several years, many researches have studied the notion of local density in order to better understand the properties of the structure of real-world networks leading especially to applications like recommendation and link prediction systems [24], [21], [22]. A better understanding of the structure of bipartite networks should help contributing to this area.

Acknowledgment

This work is supported in part by the French National Research Agency contract CODDDE ANR-13-CORD-0017-01.

References

[1] F. Le Fessant, S. Handurukande, A.-M. Kermarrec, and L. Massoulié, "Clustering in peer-to-peer file sharing workloads," in *Peer-to-Peer Systems III*. Springer, 2005, pp. 217–226.

[2] C. Prieur, D. Cardon, J.-S. Beuscart, N. Pissard, and P. Pons, "The stength of weak cooperation: A case study on flickr," *arXiv preprint arXiv:0802.2317*, 2008.

[3] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[4] A. Vzquez, R. Dobrin, D. Sergi, J.-P. Eckmann, Z. N. Oltvai, and A.-L. Barabsi, "The topological relationship between the large-scale attributes and local interaction patterns of complex networks," *Proceedings of the National Academy of Sciences*, vol. 101, no. 52, pp. 17 940–17 945, 2004.

[5] M. E. Newman, D. J. Watts, and S. H. Strogatz, "Random graph models of social networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. Suppl 1, pp. 2566–2572, 2002.

[6] F. Tarissan and R. Nollez-Goldbach, "The network of the international criminal court decisions as a complex system," in *ISCS 2013: Interdisciplinary Symposium on Complex Systems*, ser. Emergence, Complexity and Computation, A. Sanayei, I. Zelinka, and O. E. Rossler, Eds., vol. 8. Springer, 2013, pp. 225–264.

[7] R. F. i Cancho and R. V. Solé, "The small world of human language," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 268, no. 1482, pp. 2261–2265, 2001.

[8] S. Battiston and M. Catanzaro, "Statistical properties of corporate board and director networks," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 38, no. 2, pp. 345–352, 2004.

[9] M. E. Newman, S. H. Strogatz, and D. J. Watts, "Random graphs with arbitrary degree distributions," *Physics Reviews E*, vol. 64, 2001.

[10] M. Latapy, C. Magnien, and N. D. Vecchio, "Basic notions for the analysis of large two-mode networks," *Social Networks*, vol. 30, no. 1, pp. 31–48, 2008.

[11] M. Tumminello, S. Miccichè, F. Lillo, J. Piilo, and R. N. Mantegna, "Statistically validated networks in bipartite complex systems," *PloS one*, vol. 6, no. 3, p. e17994, 2011.

[12] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási, "Flavor network and the principles of food pairing," *Scientific reports*, vol. 1, 2011.

[13] J.-L. Guillaume and M. Latapy, "Bipartite graphs as models of complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 371, no. 2, pp. 795–813, 2006.

[14] F. Tarissan, B. Quoitin, P. Mérindol, B. Donnet, J.-J. Pansiot, and M. Latapy, "Towards a bipartite graph modeling of the internet topology," *Computer Networks*, vol. 57, no. 11, pp. 2331–2347, 2013.

[15] F. Tarissan, "Comparing overlapping properties of real bipartite networks," in *ISCS 2014: Interdisciplinary Symposium on Complex Systems*, ser. Emergence, Complexity and Computation, A. Sanayei, I. Zelinka, and O. E. Rossler, Eds., vol. 14. Springer, 2014, pp. 309–318.

[16] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," *Knowledge and Information Systems*, vol. 42, no. 1, pp. 181–213, 2015.

[17] K. Emamy and R. Cameron, "CiteULike: A researcher's social bookmarking service," *Ariadne*, no. 51, 2007.

[18] T. Althoff, D. Borth, J. Hees, and A. Dengel, "Analysis and forecasting of trending topics in online media streams," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 907–916. [Online]. Available: http://doi.acm.org/10.1145/2502081.2502117

[19] T. Opsahl, "Triadic closure in two-mode networks: Redefining the global and local clustering coefficients," *Social Networks*, vol. 35, no. 2, pp. 159–167, 2013.

[20] P. Zhang, J. Wang, X. Li, M. Li, Z. Di, and Y. Fan, "Clustering coefficient and community structure of bipartite networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 27, pp. 6869–6875, 2008.

[21] S. P. Borgatti and M. G. Everett, "Network analysis of 2-mode data," *Social networks*, vol. 19, no. 3, pp. 243–269, 1997.

[22] T. A. Snijders, "The statistical evaluation of social network dynamics," *Sociological methodology*, vol. 31, no. 1, pp. 361–395, 2001.

[23] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[24] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.