Tag Spam Creates Large Non-Giant Connected Components

Nicolas Neubauer
Neural Information
Processing Group
Technische Universität Berlin
neubauer@cs.tu-
berlin.de

Robert Wetzker
DAI Labor Technische
Universität Berlin
robert.wetzker@dai-
labor.de

Klaus Obermayer
Neural Information
Processing Group
Technische Universität Berlin
oby@cs.tu-berlin.de

ABSTRACT

Spammers in social bookmarking systems try to mimick bookmarking behaviour of real users to gain the attention of other users or search engines. Several methods have been proposed for the detection of such spam, including domain-specific features (like URL terms) or similarity of users to previously identified spammers. However, as shown in our previous work, it is possible to identify a large fraction of spam users based on purely structural features. The hypergraph connecting documents, users, and tags can be decomposed into connected components, and any large, but non-giant components turned out to be almost entirely inhabitated by spam users in the examined dataset. Here, we test to what degree the decomposition of the complete hypergraph is really necessary, examining the component structure of the induced user/document and user/tag graphs. While the user/tag graph's connectivity does not help in classifying spammers, the user/document graph's connectivity is already highly informative. It can however be augmented with connectivity information from the hypergraph. In our view, spam detection based on structural features, like the one proposed here, requires complex adaptation strategies from spammers and may complement other, more traditional detection approaches.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning; G.2.2 [**Graph Theory**]:

General Terms

Tagging, Connected Components, Spam Detection

1. INTRODUCTION

Social bookmarking systems by now have been drawing enough attention to create incentives for spammers to pol-

lute their data. In order to promote a website, they pretend to be saving it for later reference like normal users. As those posts are typically publicly visible, they hope to either trick users into visiting their website or to be rewarded by search engines for being linked from a high-profile website such as the social bookmarking site. There is a growing body of research on this topic from simulating the impact of spam users on the overall dataset [3] over the analysis of suspicious patterns in an unlabelled dataset[11] to actual prediction approaches. These approaches are typically based either on domain-specific features of the posted items, or on methods that create similarities between users such that evidence can be propagated from known to unknown users ([5],[2],[7]).

This article extends previous work[8] on connectivity in tagging datasets, further exploring the role of spam behaviour under various definitions of connectivity. We presuppose there are fundamental differences between legitimate and spamming bookmarking behaviour, and that these differences should be mirrored in the structure of the resulting data. One such structural property is the distribution of connected components, those subgraphs of a graph which do not share connections among each other. In the following, we define three different ways for the creation of connected components of graphs derived from the original data. We apply our aproach to the Bibsonomy social bookmarking dataset and find a salient giant component and spam-polluted next-largest components in both the complete hypergraph and the user/document graph, i.e., ignoring tags. Based on these findings, we propose a simple user spam predictor based on membership in the giant or the next-largest components of the two graphs. We can show that the complete hypergraph's and the user/document graph's connectivity structures contain at least partially complementary information. We conclude by discussing the pros and cons of the proposed approach and listing possible future extensions.

2. CONNECTED COMPONENTS

Each single event of a user u tagging a document d with a tag t can be interpreted as an edge connecting the three nodes (d, u, t). The set of all edges then defines a 3-partite (because connected elements are from three different sets) 3-uniform (because each edge connects exactly three nodes) hypergraph H. Interpreting social bookmarking data as a graph structure allows us to apply a basic analytic tool for complex graphs: the examination of connected components. The connected components of a graph define its disjunct subgraphs, i.e. a partition of its nodes such that a path ex-

(a) H with one component   (b) H and its 2-hyperincident connected components
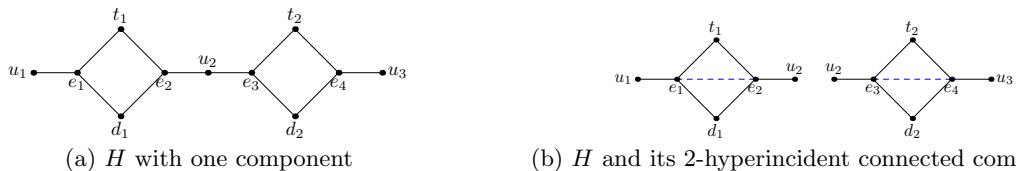
Figure 1: A sample 3-uniform hypergraph with edges plotted as nodes e_i. The blue, dotted lines indicate 2-hyperincidence between edges e_1 and e_2 / e_3 and e_4, respectively. e_2 and e_3 are incident via u_2, but not 2-hyperincident, turning a single connected component into two 2-hyperincident connected components

ists between any pair of nodes within a component, but no path exists between any two nodes from two different components. The size distribution of connected components and in particular the existence and relative size of a so-called "giant component", i.e., a single connected subgraph containing a majority of nodes, is a very well-researched phenomenon that can yield valuable insights into the underlying formation dynamics. For example, a graph created by randomly adding edges to a fixed set of vertices almost certainly exhibits a so-called "percolation transition", i.e., the emergence of a giant connected component, when the ratio of edges to vertices exceeds 0.5[1].

Decomposing tagging networks into their connected components using the classic notion of connectivity, however, turns out to be uninformative: they tend to be almost entirely connected. Therefore, we propose three alternatives.

The *user/document-graph UD(H)* is defined by the edges $\{(d, u) : \exists (d, u, t) \in H\}$, i.e., tags are ignored and only shared documents imply connectedness.

The *user/tag-graph UT(H)* analogously is defined by the edges $\{(u, t) : \exists (d, u, t) \in H\}$.

Hyperincident-connected components[8] allow for the direct decomposition of H in spite of its high connectivity. We basically raise the criterion for being connected: We say two edges are m-hyperincident if they share not one, but m nodes. Then, m-hyperincident components can be defined as partitions of edges such that paths of m-hyperincident edges exist between all members of a component, but not between members from two different components. Since this definition partitions edges instead of nodes, nodes can be part of several connected components. Figure 1 shows an example of a 3-hypergraph and its 2-hyperincident connected components.

To see why such a definition might be useful, consider a spammer who tries to appear legitimate by tagging 'cnn.com' with 'news' (assuming this is a frequent association). The corresponding edge ('cnn.com', spamuser, 'news') would be connected to the giant component, since the new edge is 2-incident to the assumed edges ('cnn.com', realuser, 'news') and correctly be considered legitimate. Consider a second entry by that user: ('spam.com', spamuser, 'spamtag'). Assuming that neither 'spam.com' nor 'spamtag' are part of the giant component, this edge is incident, via the user, but not 2-incident to any edge in the giant component. So while normal connectivity conditions would would join this edge to the giant component, the stricter conditions automatically counteract basic cloaking measures of spammers and keep such entries isolated. We will now examine how much additional information we gain by this in practice.

3. ANALYSIS

Analyses were performed on the Bibsonomy dataset[9] as provided to the participants of the PKDD/ECML 2008 Tag Spam Discovery Challenge[4]. It consists of 16,818,699 edges connecting 1,574,963 documents, 396,474 tags and 38,920 users. 93% of all users have been hand-labelled as spammers. In the following, we present an analysis on this dataset in its complete form (black) and on a cleaned version containing only non-spam data (green).

Figure 2 presents the distribution of component sizes of the different induced graphs. Figure 3 details the sizes of the ten largest components of each. First of all, we see that $UT(H)$ is almost entirely connected both in the spam and the non-spam dataset. For $UD(H)$ and H, however, we see a distinct giant component followed by much smaller (around two orders of magnitude) components in the non-spam dataset. This distinction, however, is weakened in the spam dataset. It turns out that only 82.5%/80.6% of the users in $UD(H)$'s / H's giant component are spammers (compared to 93% in the overall population), whereas the following next-largest components are made up almost exclusively by spammers. Figure 4 quantifies this notion by plotting for each class (spam/non-spam user) the fraction of its members in the different types of components. Here, we can see in particular that the fraction of legitimate users in next-largest components is marginal.

The conclusion from these graphs is that spam behaviour creates groups of users tagging documents which the majority of legitimate users is not interested in. The decomposition of $UT(H)$ does not show such behaviour, which might imply this separation is based on document selection entirely. However, we see the connectivity patterns slightly differ in the hypergraph. This implies tags do play a subtle role which however needs to be exploited in combination with the corresponding documents.

4. SPAM PREDICTION

Based on the results of Figure 4 and to further explore the relation between $UD(H)$'s and H's connectivity, we devise a simple classification scheme. Users are classified as non-spam if contained in the giant component. If isolated, we judge them neutrally, and membership in a large but non-giant component leads to the user being labeled as a spammer. We combine (blue) the predictions based on $UD(H)$ (red) and the hypergraph (black) such that users rated as spammers via both graphs receive a higher spam rating than those only rated as spam by one. Since we assume that many of the isolated components are caused by users which have tagged only very few documents, we additionally evaluate these heuristics on the subsets of users having tagged more
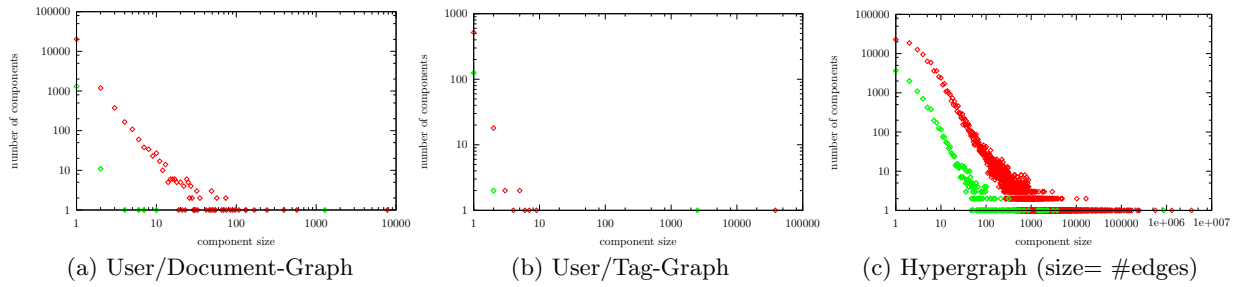
(a) User/Document-Graph   (b) User/Tag-Graph   (c) Hypergraph (size= #edges)

Figure 2: Distribution of component sizes – component size (x) vs. number of components of that size (y). Distributions roughly follow a power law-like shape in both spam (black) and non-spam (red) datasets.

(a) User/Document-Graph   (b) User/Tag-Graph   (c) Hypergraph (edges/component)
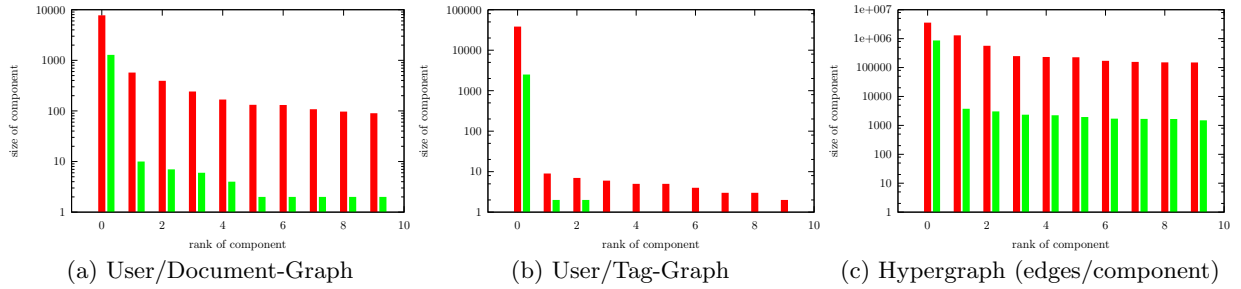
Figure 3: Size of 10 largest components – the size of the next-largest components decays sharply in the non-spam (green) dataset (several orders of magnitude), but much more smoothly when spam (black) is included.

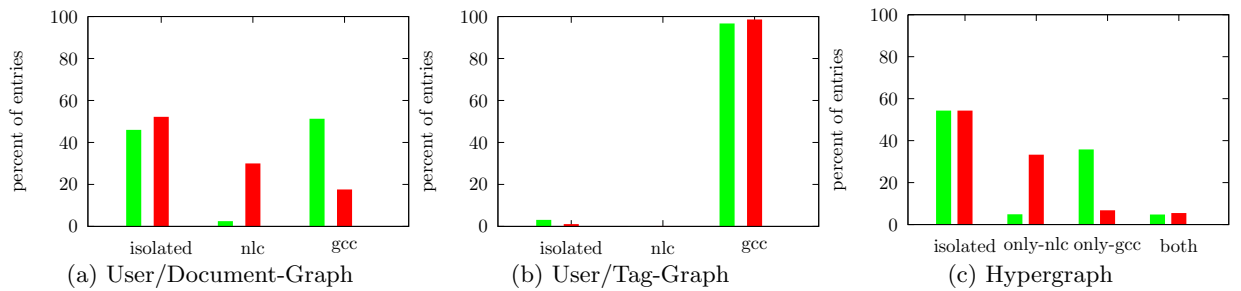(a) User/Document-Graph   (b) User/Tag-Graph   (c) Hypergraph

Figure 4: Distribution of spam (black) and non-spam (green) users over different classes: Isolated components containing only a single user, next-largest components (nlc) containing more than one user, but not being the giant component, or the giant component. In the case of the hypergraph, users can be part of both the giant and next-largest components (since component membership is unique only for edges.)

(a) all users   (b) users $\geq$ one document   (c) users $\geq$ 10 documents   (d) users $\geq$ 100 documents
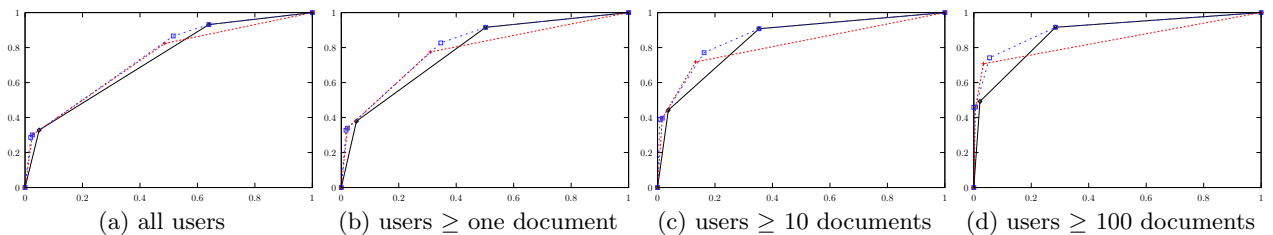
Figure 5: ROC curves for user spam prediction based on component membership in the user/document graph (red), the entire hypergraph (black) or a combination of both (blue). Curves are produced by plotting the ratio of false positives (x) against the ratio of true positives (y) as the classification threshold decreases from maximum to minimum. Discrete values are assigned based on membership (see text), producing edges in the ROC curves whenever the threshold passed one of those values.

than 1, 10, or 100 documents (Figures 4 b), c) and d)).

Table 1 shows the performance values when applying these heuristics. The employed AUC measure (area under the curves displayed in Figure 5) is a balanced accuracy measure taking into account that the class of spam users is much bigger: Labelling each user as a spammer would create an accuracy of 93%, but an AUC of exactly 0.5%. As was to be expected from the distribution of users, $UT(H)$ does not provide any usable information. The other approaches however do well, particularly on users with more documents. The hypergraph's connectivity is slightly more informative than $UD(H)$'s. However, the performance of the combined predictor exceeds that of both individual ones. This suggests that neither graph's connectivity is redundantly encoded by the other's, i.e., there must be users which are in a next-largest component in one graph but not in the other.

Figure 5 further explores the relation between the three classifiers. We see the ROC curves for each classifier and each subset of users. It turns out that $UD(H)$ provides slightly less false positives for the spam condition (producing the first slope), whereas the hypergraph is more expressive for the non-spam condition (responsible for the third slope). The simple combination of the classifiers unites both advantages, as can be seen by its ROC curve being an almost perfect upper bound for the other two's.

5. DISCUSSION

We have examined different decompositions of a tagging dataset into connected components. The results suggest that the characteristic giant component of the entire hypergraph, also found in other tagging datasets[8], is largely caused by user/document connections, which mirror this distribution. Although $UT(H)$ does not show any meaningful component structure, the full hypergraph's connectivity contains additional structure not present in $UD(H)$. This suggests that tags can contain meaningful connectivity information, but more subtly than what could be captured by $UT(H)$. This was also implied by the improved performance of the combined classification heuristics.

The introduced heuristics cannot reach the performance of highly specialized classifiers. Nevertheless, they might be used as building blocks of more complex classifiers. Independent from content or prior labels, they can also be used under a wider range of conditions – for example, when the tagged resources are more difficult to extract features from than URLs, e.g. movies. More generally, we believe identifying behavioural patterns will help creating spam prediction mechanisms that are harder to fool. In bookmarking their favourite resources, human beings create traces of complex cognitive processes. Finding properties that separate data resulting from those processes from that of shallow spamming activities should considerably raise the costs of creating innocent-looking spam.

Nevertheless, our decompositions remain somewhat coarse particularly for users with few documents. Besides validating our findings on additional datasets with spam, we want to explore possibilites of further restricting connectivity in the future. In [10], the authors explore the connectivity of graphs between users only, created through thresholded similarity functions. These and other aspects of connectivity, like the temporal evolution of component structure[6], can yield further insight into the dynamics of legitimate and spamming bookmarking behaviour.

	min # docs/user			
	0	1	10	100
User/Document	0.73	0.78	0.81	0.84
User/Tag	0.49	0.49	0.50	0.50
Hypergraph	0.73	0.78	0.84	0.88
Combined	0.76	0.81	0.87	0.91

Table 1: AUC values of the different connectivity-based classifiers

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] P. Erdos and A. Renyi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.

[2] A. Gkanogiannis and T. Kalamboukis. A novel supervised learning algorithm and its use for spam detection in social bookmarking systems. In *ECML PKDD Discovery Challenge 2008 (RSDC'08)*, 2008.

[3] P. Heymann, G. Koutrika, and H. Garcia-Molina. Fighting spam on social web sites: A survey of approaches and future challenges. *IEEE Internet Computing*, 11(6):36–45, 2007.

[4] A. Hotho, D. Benz, R. Jäschke, and B. Krause, editors. *ECML PKDD Discovery Challenge 2008 (RSDC'08)*. Workshop at 18th Europ. Conf. on Machine Learning (ECML'08) / 11th Europ. Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'08), 2008.

[5] B. Krause, A. Hotho, and G. Stumme. The anti-social tagger - detecting spam in social bookmarking systems. In *Proc. of the Fourth International Workshop on Adversarial Information Retrieval on the Web*, 2008.

[6] M. McGlohon, L. Akoglu, and C. Faloutsos. Weighted graphs and disconnected components: patterns and a generator. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 524–532, New York, NY, USA, 2008. ACM.

[7] N. Neubauer and K. Obermayer. Predicting tag spam examining cooccurrences, network structures and url components. In *ECML PKDD Discovery Challenge 2008 (RSDC'08)*, 2008.

[8] N. Neubauer and K. Obermayer. Hyperincident components of tagging networks (submitted). In *HyperText 2009, Proceedings of*, 2009.

[9] Knowledge Discovery and Data Engineering Group, University of Kassel. Benchmark folksonomy data from bibsonomy, version of june 30th, 2008.

[10] E. Santos-Neto, M. Ripeanu, and A. Iamnitchi. Tracking usage in collaborative tagging communities.

[11] R. Wetzker, C. Zimmermann, and C. Bauckhage. Analyzing social bookmarking systems: A del.icio.us cookbook. In *Mining Social Data (MSoDa) Workshop Proceedings, ECAI 2008*, pages 26–30, 2008.