Connected Components of 3-Partite 3-Uniform Hypergraphs

Nicolas Neubauer[1] and Klaus Obermayer
Neural Information Processing Group, Technische Universität Berlin
ni.cs.tu-berlin.de

Social bookmarking sites allow users to organize resources like URLs, bibliographic references or media files using "tags", arbitrary keywords that can be used later on to retrieve those resources. As many users (millions, on the most popular sites) perform such tagging in parallel, this leads to a rich, collaborative description of the tagged resources[2]. The resulting data can be interpreted as a 3-partite 3-uniform graph formed by edges (d, u, t) for each document d being tagged by user u with tag t, opening the toolbox of complex network analysis for their study. We study the role of the connected components[1] of the hypergraph and derived structures. In particular, we try to tell apart legitimate tagging activity from that of "tag spammers" which use social bookmarking systems to create undesired references to the advertised website[3].

Decomposing the hypergraph into connected components is uninformative, since it basically consists of a single, connected component. Therefore, we compare three derived connectivity measures: First, we use a generalized notion of connectivity on hypergraphs, hyperincident connectivity[4], which requires edges to share more than one node. Second, we reduce the hypergraph to bipartite graphs by ignoring either documents, users, or tags, such that, for example, when ignoring tags, (d, u) is an edge of the reduced graph for all $(d, u) : (d, u, t) \in H$. Third, we examine the induced graphs for given documents, users, or tags, such that (d, u), for example, is an edge of the induced graph of tag t iff $(d, u, t) \in H$.

We find that genuine human activity in all of these cases creates a characteristic connectivity structure typically involving a salient giant component and much smaller next-largest components. This pattern is destroyed by the fake tagging activity of spam users: Complex next-largest components are created that in many cases are entirely made up of spam. We believe such an unsupervised examination is valuable not only for creating more efficient and harder to fool anti-spam measures, but also for deepening the understanding of the collective cognitive processes underlying the creation of such networks.

Preference: Poster

Topic: Data Mining

References

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

[2] S. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, April 2006.

[3] 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.

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

[1] presenting author