The Blogosphere at a Glance—Content-Based Structures Made Simple

Olof Görnerup[1] and **Magnus Boman**[1,2]
[1]Swedish Institute of Computer Science (SICS), SE-164 29 Kista, Sweden
[2]Royal Institute of Technology (KTH/ICT/SCS), SE-164 40 Kista, Sweden
{olofg, mab}@sics.se

Abstract

A network representation based on a basic word-overlap similarity measure between blogs is introduced. The simplicity of the representation renders it computationally tractable, transparent and insensitive to representation-dependent artifacts. Using Swedish blog data, we demonstrate that the representation, in spite of its simplicity, manages to capture important structural properties of the content in the blogosphere. First, blogs that treat similar subjects are organized in distinct network clusters. Second, the network is hierarchically organized as clusters in turn form higher-order clusters: a compound structure reminiscent of a blog taxonomy.

1 Introduction

Several tools and algorithms have been developed for harnessing the vast amount of data that constitutes the blogosphere (cf. [Agarwal and Liu, 2008; 2009]), e.g., by collecting, relating and visualizing blog entries [Tauro *et al.*, 2008; Llor *et al.*, 2007; Uchida *et al.*, 2007; Bross *et al.*, 2010] or tag clouds, [Fujimura *et al.*, 2008], or by classifying blogs in terms of interblog communication and community stability [Chi *et al.*, 2007], sense of community among bloggers [Chin and Chignell, 2006], discussion keyword correlation [Bansal *et al.*, 2007], and a host of machine learning and statistics approaches, cf. [Tsai, 2011]. To date, however, almost all these tools and algorithms require human intervention and considerable time investment to overcome problems with bootstrapping, tuning, and not least semantics. Understanding a graph, perhaps with thousands of vertices and edges, pertaining to describe relevance to one's own blog according to some set of possibly esoteric or advanced criteria is not straightforward. We address this problem by presenting a method for generating a network of relevant blogs by means of the simplest similarity criterion there is: word overlap. We will demonstrate that even this naïve approach allows us to capture fundamental and important structural properties of the blogosphere.

2 Method

We represent the blogosphere as a network, where nodes constitute blogs, and where blogs are linked if they have similar textual content. Links are weighted, where the strength of a link is given by a similarity measure.

2.1 Similarity measure

To estimate the similarity between blogs we simply compare the overlap of occurring words. Given two blogs i and j, let $\mathcal{W}_i$ denote a set of words (to be specified below) that occur in i, and $\mathcal{W}_j$ a set of words that are used in j. The similarity s_{ij} between i and j is then defined as the Jaccard index

$$s_{ij} = \frac{|\mathcal{W}_i \cap \mathcal{W}_j|}{|\mathcal{W}_i \cup \mathcal{W}_j|}. \quad (1)$$

In other words, s_{ij} is the fraction of all words in $\mathcal{W}_i$ and $\mathcal{W}_j$ that are shared by the two sets. It holds that $0 \leq s_{ij} \leq 1$, where $s_{ij} = 1$ if $\mathcal{W}_i$ and $\mathcal{W}_j$ are identical and $s_{ij} = 0$ if they do not share a single word. This similarity measure is equivalent to Tversky's Ratio model [Tversky, 1977], which has been found to be a good trade-off between simplicity and performance among text document similarity measures [Lee *et al.*, 2005].

2.2 Word filtering

We do not consider the full word sets of blogs—literally *all* occurring words—for several reasons. Comparing very common words ("the", "it", "do", etc.) will only provide a negligible amount of similarity information. The use of uncommon words, on the other hand, is likely to tell us a lot about the characteristics of a blog. However, at the same time we do not want to consider words that are too uncommon—for instance those occurring only a handful of times in the blogosphere during the course of several months—since these are often misspellings and typos that only add noise to the statistics. Another reason for not considering all words is a pragmatic one. Analyzing tens of thousands of blogs can be computationally expensive. By utilizing Zipf's law [Zipf, 1949], which implies that a few of the most common words represent a large majority of word occurrences[1], the computational cost is drastically reduced.

[1]More specifically, the frequency of a word is inversely proportional to its rank; $f_n \sim 1/n^a$, where n is the rank ($n = 1$ for the most common word, $n = 2$ for the second most common word, etc.) and a is some exponent.

2.3 Network structure

The global structure of a similarity network may provide valuable information about how blogs and groups of blogs are related with respect to contents. We have focused on two network properties: Community structure and hierarchical organization.

Complex networks typically exhibit communities, where nodes are clustered in groups [Newman, 2003]. Characteristic for community structures is that there are significantly higher densities of edges within communities than between them. This property may be quantified as follows [Newman and Girvan, 2004]: Let $\{v_1, v_2, ..., v_n\}$ be a partition of a set of vertices into n groups, r_i the degree of edge weights (i.e., similarities) internal to v_i (the sum of internal weights over the sum of all weights in the network) and s_i the degree of weights of edges that start in v_i. The degree of community structure is then defined as

$$Q = \sum_{i=1}^{n} (r_i - s_i^2). \quad (2)$$

To infer clusters in the blog network we have employed an agglomerative clustering technique [Clauset, 2005], that aims to find cluster assignments—a partition of the set of vertices—that maximizes the community structure measure Q.

Another method by Clauset *et al*. [Clauset *et al.*, 2008] has been used to identify the hierarchical structure of the blog network. This method combines a maximum likelihood approach with a Monte Carlo sampling procedure to infer likely hierarchical models of the network.

2.4 Case study: The Swedish blogosphere

We have tested our approach on the Swedish blogosphere. The API of the blog search engine *Twingly*[2] has been used for collecting blog posts from a five-month period. The posts were fetched and aggregated (i.e., for each blog, posts were concatenated). In the spirit of keeping things simple we refrained from applying *ad hoc* textbook pre-processing such as stemming and relied on basic word frequency statistics to filter out words: First we discarded all words occurring less than ten times. Of the remaining words we then kept those that occurred in the fifth percentile of the frequency distribution. For each blog, we collected its set of those occurring words. Blogs that had word sets of size 25 or larger were kept. This ensured a meaningful similarity measure and also filtered out a considerable amount of spam blogs. At this point, 21564 blogs remained. We have varied the above parameters in sensitivity analyses, and the results reported here appear to be stable.

3 Results

The content-based blog network is found to have a distinct clustered structure. We have visualized this by plotting edges with a weight above a certain threshold (i.e. only relations between highly similar blogs are shown) such that blog communities crystallize into separate subnetworks. See Fig. 1, where we plot the acquired network with various weight thresholds. By inferring communities and then inspecting the actual content of blogs within communities, we find that the clusters reflect topics domains such as politics, books, technology, or music, cf. Fig. 2. Note that spam blogs, *splogs*, also form separate clusters. Splogs are in fact particularly tightly knit, presumably since they tend to contain homogenous sets of words.

Furthermore, when employing Clauset *et al*.'s hierarchy inference algorithm, we find that clusters indeed are organized in higher order (meta-) clusters. An example of the hierarchical organization of the blog network is depicted in Fig. 3 in the form of a consensus dendrogram—i.e., a dendrogram that is consistent with several inferred hierarchical models—of a "food and beverages" cluster. There we see that food and beverages are separated into two clusters, and the beverage cluster in turn consists of a wine and a beer cluster. Again, the validity of acquired hierarchies is evaluated by inspection.

4 Discussion and outlook

We have shown that the signal in raw blog data is so strong that even our basic similarity measure—word occurrence overlap—is capable of capturing valuable structural information. The measure is computationally tractable and enables efficient categorization of blogs when used in concurrence with fast graph clustering algorithms. We grant that there are more advanced—and possibly more accurate—(document) similarity measures [Agarwal *et al.*, 2008; Elsas *et al.*, 2008; Lee *et al.*, 2005; Macdonald and Ounis, 2008]. However, we believe that the minimal (non-trivial) measure employed here is suitable as a baseline when studying blog similarity networks. The measure is admittedly simplistic, yet this is also its strength since it decreases the risk of causing hidden representation-dependent artifacts that are more difficult to identify when using more advanced similarity measures.

Because of the rapid growth of data in the blogosphere, there is a strong demand from industry as well as from research for simple means to harvesting blog data. Our approach is obviously among the simplest possible, but we have not discussed any explict applications here, since employment is not our chief concern. Neither have we provided any analyses of computational complexity, because such analyses will be application-driven and will likely contain very detailed average-case, rather than general worst-case, complexity measures.

An issue that needs to be addressed in future work is that of validation. How can we know that acquired blog clusters are meaningful? So far, our approach has been to examine a random sample of blogs and subjectively confirm that their contents is consistent within inferred blog clusters. Such empirical evaluations can be problematic, however. In some cases, a manual classification may be considered as clear cut (e.g., identifying that two blogs that solely treat Belgian beer belong to the same cluster), but not always. A more quantitative measure that validates the result is therefore desirable. This can, on the other hand, also be turned into an epistemological question. One can for example imagine cases when the blog classes acquired from the similarity network can be used to evaluate *other* blog classifications (including our own subjec-

[2] http://www.twingly.com/

tive one). However, in this discussion we have more pragmatic and application-oriented evaluation methods in mind.

We have treated only a few structural aspects of the blog network here. These deserve more attention, as do the dynamics and evolution of the networks: How does information diffuse and change in the network, and how does the network structure itself change over time? For instance, through an analysis along these lines one may perhaps trace how emerging trends or news proliferate in and between specific topic domains of the blog similarity network.

Another possible future direction concerns splog detection. We have observed that splogs emerge as separate categories. If an individual blog is identified as a splog (e.g., by examining the distribution of blog similarities), it is likely that its associated blog cluster also consists of splogs. If such a relation proves to hold true in general, it enables splog detection and removal at the level of blog clusters rather than individual blogs, which presumably would be much more efficient.

As the network representation of the blogosphere is found to be hierarchically structured, it may pave the way for applications that operate on different levels of resolution; from blogs to groups of similar blogs, to groups of groups of blogs, and so forth. The hierarchical organization also enables a top down approach to blog navigation that starts at a coarse level of blog categories and then narrows down to finer scales. Monitoring may also be more efficient and accurate if limited to a specific and relevant topic-domain of blogs. That is, although the blogosphere may seem overwhelming at times, it is in fact intrinsically structured in terms of content as to enable effective navigation and monitoring.

Acknowledgements

OG was funded by The Internet Infrastructure Foundation (.SE). The authors thank Twingly for providing blog data and Aaron Clauset for sharing source code for the hierarchical structure inference algorithm and for the radial dendrogram visualization script used for rendering Fig. 3. The authors also thank Jussi Karlgren for providing some of the references to earlier work.

References

[Agarwal and Liu, 2008] Nitin Agarwal and Huan Liu. Blogosphere—research issues, tools, and applications. *SIGKDD Explorations*, 10(1):18–31, 2008.

[Agarwal and Liu, 2009] Nitin Agarwal and Huan Liu. *Modeling and Data Mining in Blogosphere*. Morgan and Claypool Publishers, 2009.

[Agarwal *et al.*, 2008] Nitin Agarwal, Huan Liu, Lei Tang, and Philip S. Yu. *Identifying the Influential Bloggers in a Community*. In *Proceedings of the international conference on Web search and web data mining*. ACM, New York, 2008.

[Bansal *et al.*, 2007] Nilesh Bansal, Nick Koudas, Fei Chiang, and Frank Wm. Tompa. Seeking stable clusters in the blogosphere. In *Proceedings of the 33rd international conference on Very large data bases*, pages 806–817. VLDB Endowment, 2007.

[Bross *et al.*, 2010] Justus Bross, Matthias Quasthoff, Philipp Berger, Patrick Hennig and Christoph Meinel Mapping the Blogosphere with RSS-Feeds. In *Proceedings of the 2010 24th IEEE International Conference on Advanced Information Networking and Applications*, pages 453–460. IEEE Computer Society, Washington, DC, 2010.

[Chi *et al.*, 2007] Yun Chi, Shenghuo Zhu, Xiaodan Song, Junichi Tatemura, and Belle Tseng. Structural and temporal analysis of the blogosphere through community factorization. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 163–172. ACM, New York, 2007.

[Chin and Chignell, 2006] Alvin Chin and Mark Chignell. A social hypertext model for finding community in blogs. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 11–22. ACM, New York, 2006.

[Clauset *et al.*, 2008] A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, 2008.

[Clauset, 2005] Aaron Clauset. Finding local community structure in networks. *Physical Review E*, 72:026132, 2005.

[Elsas *et al.*, 2008] Jonathan L. Elsas, Jaime Arguello, Jamie Callan, and Jaime G. Carbonell. *Retrieval and feedback models for blog feed search*. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, 2008.

[Fujimura *et al.*, 2008] Ko Fujimura, Shigeru Fujimura, Tatsushi Matsubayashi, Takeshi Yamada and Hidenori Okuda. Topigraphy: visualization for large-scale tag clouds. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 1087–1088. ACM, New York, 2008.

[Lee *et al.*, 2005] Michael D. Lee, Brandon Pincombe, and Matthew Welsh. An empirical evaluation of models of text document similarity. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1254–1259. Erlbaum, 2005.

[Llor *et al.*, 2007] Xavier Llor, Noriko Imafuji Yasui, Michael Welge, and David E. Goldberg. Human-centered analysis and visualization tools for the blogosphere. In *Proceedings of the Digital Humanities 2007*, 2007.

[Macdonald and Ounis, 2008] Craig Macdonald and Iadh Ounis. Key blog distillation: ranking aggregates. In *Proceedings of the 17th ACM Conference on Information and knowledge management*. ACM, New York, 2008.

[Newman, 2003] Mark Newman. The Structure and Function of Complex Networks. *SIAM Review*, 2:167–256, 2003.

[Newman and Girvan, 2004] Mark Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 2004.

[Tauro *et al.*, 2008] Candida Tauro, Sameer Ahuja, Manuel A. Prez-Quiones, Andrea Kavanaugh, and Philip Isenhour. Vizblog: Discovering conversations in the blogosphere. In *Technology demonstration at Directions and Implications of Advanced Computing - Conference on Online Deliberation*, University of California, Berkeley, 2008.

[Tsai, 2011] Flora S. Tsai. Dimensionality reduction techniques for blog visualization. *Expert Systems with Applications*, 38(3):2766–2773, 2011.

[Tversky, 1977] Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.

[Uchida *et al.*, 2007] Makoto Uchida, Naoki Shibata, and Susumu Shirayama. Identification and visualization of emerging trends from blogosphere. In *Proceedings of International Conference on Weblogs and Social Meida*, pages 305–306, 2007.

[Zipf, 1949] George K. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison Wesley, Cambridge MA, 1949.

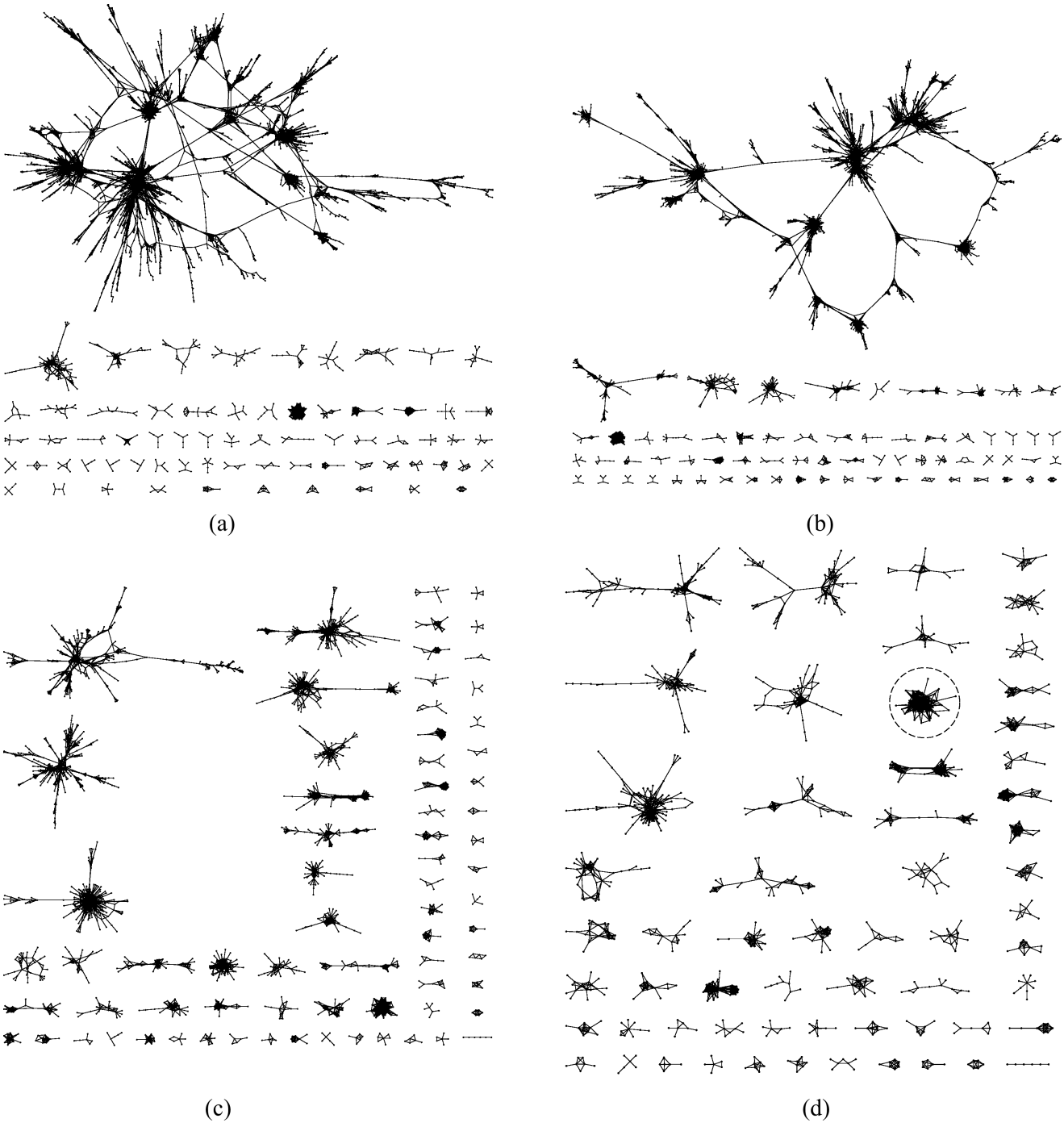

Figure 1: Visualization of the Swedish blogosphere, where blogs with similarities $\geq \gamma$ are shown. (a) $\gamma = 0.04$. (b) $\gamma = 0.045$. (c) $\gamma = 0.055$. (d) $\gamma = 0.07$. A spam blog cluster is enclosed within a dashed circle.

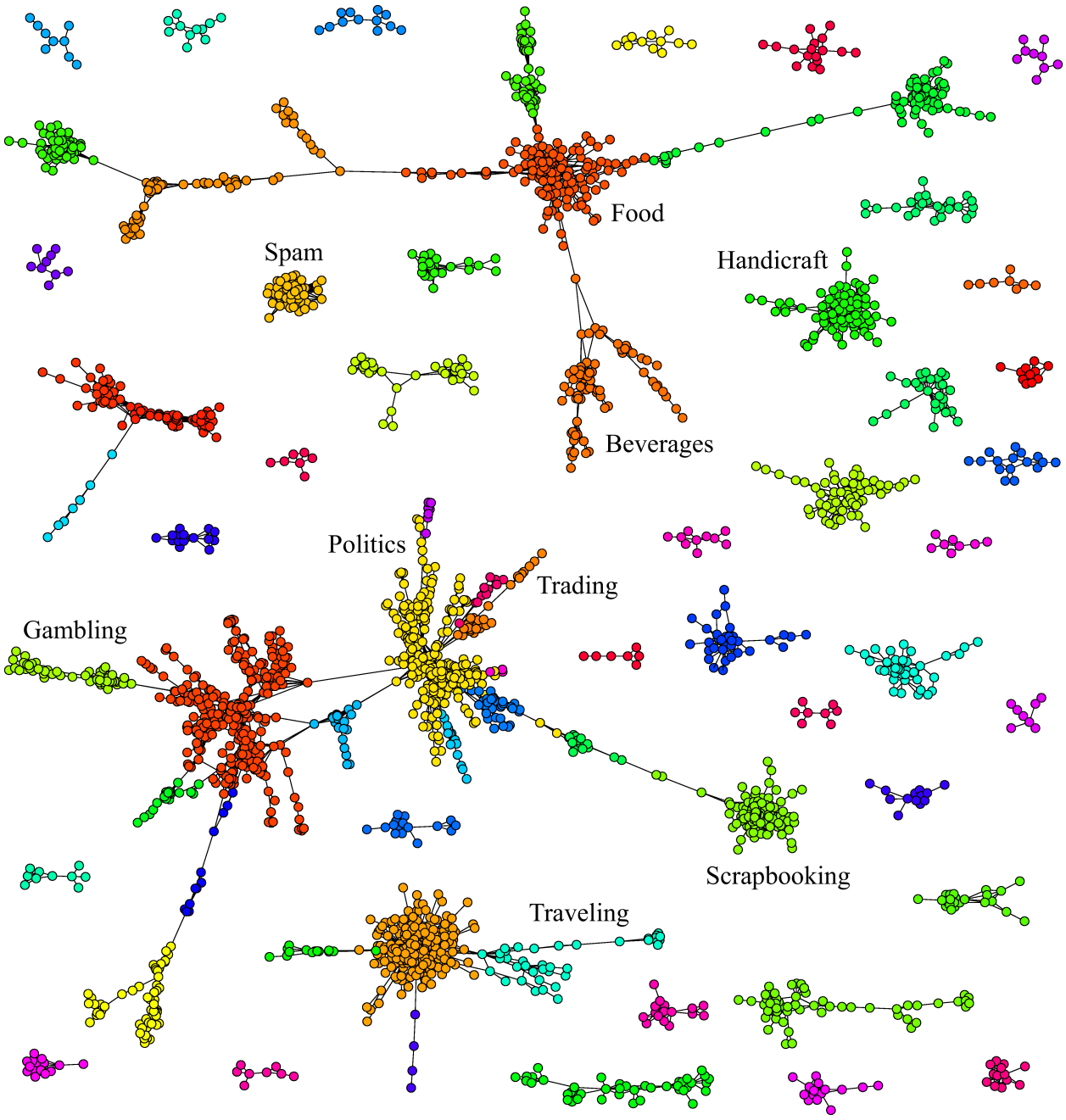

Figure 2: Content-based visualization of Swedish blogs. Blog categories (color-coded) are derived as network communities. Some example categories are labeled. For sake of clarity, only edges with weights larger than or equal to 0.05 are shown.

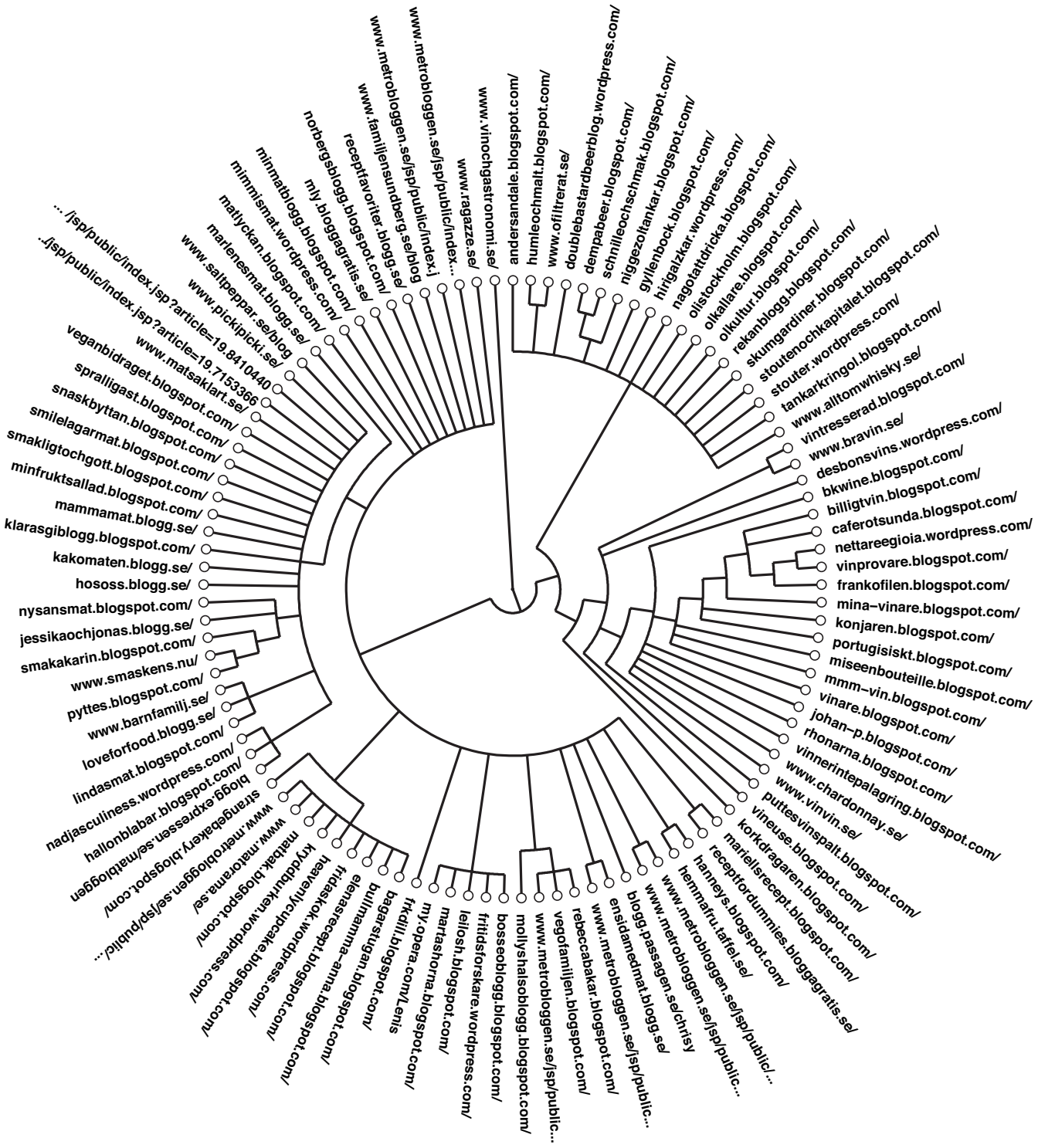

Figure 3: Consensus dendrogram of the food-and-beverages cluster, where blogs are organized in separate vine, beer and food clusters. Leaf nodes are labeled with corresponding blog URLs.