# P2P Authority Analysis for Social Communities

Josiane Xavier Parreira, Sebastian Michel, Matthias Bender,
Tom Crecelius, Gerhard Weikum
Max-Planck-Institut für Informatik
66123 Saarbrücken, Germany
{jparreir, smichel, mbender, tcrecel, weikum}@mpi-inf.mpg.de

## ABSTRACT

PageRank-style authority analyses of Web graphs are of great importance for Web mining. Such authority analyses also apply to hot "Web 2.0" applications that exhibit a natural graph structure, such as social networks (e.g., MySpace, Facebook) or tagging communities (e.g., Flickr, Del.icio.us). Finding the most trustworthy or most important authorities in such a community is a pressing need, given the huge scale and also the anonymity of social networks.

Computing global authority measures in a Peer-to-Peer (P2P) collaboration of autonomous peers is a hot research topic, in particular because of the incomplete local knowledge of the peers, which typically only know about (arbitrarily overlapping) sub-graphs of the complete graph. We demonstrate a self-organizing P2P collaboration that, based on the local sub-graphs, efficiently computes global authority scores. In hand with the loosely-coupled spirit of a P2P system, the computation is carried out in a completely asynchronous manner without any central knowledge or coordinating instance. We demonstrate the applicability of authority analyses to large-scale distributed systems.

## 1. INTRODUCTION

### 1.1 Motivation

With the recent advance of social communities, mining their structures is a relevant topic not only for research, but with immediate real-life applications. Identifying "hot spots" of a community is a vital guidance for users to identify potentially interesting entities, and captures the "Chinese whisper"-style evolution of facts and trends [9, 17, 8]. At the same time, the high growth rate of such networks together with the high dynamics makes it an increasingly hard task to separate authoritative information and players from pure rumors or sometimes even guerrilla marketing. The evolutionary development of such networks has also sparked research towards time-aware authority ranking [5].

Fortunately, the community users interact in a way that results in community graphs that allow authority analyses similar to popular PageRank-style analyses on Web graphs [6]. Such community graphs naturally arise in various applications, by different means of user interaction, with respect to a wide variety of entities, and with varying notions of authority. A popular example of such a community is the photo-sharing community Flickr, where users can upload their digital images, build friend-relationships with other users, and tag pictures (both your own and your friends'). As another representative application, consider social communities such as MySpace or Facebook, where users can upload personal profiles, maintain diary-style blogs, do real-time chatting with other users, and invite other users as "friends". There are also various examples of more traditional reference graphs, e.g., co-author or citation matrices [20]. Bibliometric analyses of such data, e.g., based on the spectral analysis of the underlying graphs, are gaining attention as a means of measuring impact.

### 1.2 Technical Issues

The notion of authority differs widely among these communities, but typically refers to a combination of factors like trustworthiness [10], reputation [14], importance or prestige of entities [15, 13], or attention level. Similar approaches have also been proposed to combat link spam [11, 16, 2].

Edges in such a graph, and thus the notion of authority transfer across entities in the graph, typically expose the same features that apply to social networks in the real world. For example, they are not necessarily symmetric (who does not know people that call you a "friend", while you would prefer to not have much to do with them...), and they show a certain degree of transitivity (if you trust your friend, you somehow also trust her friends, but after some hops you would end up trusting the whole world...).

Authority analysis on such graphs typically builds on spectral methods, i.e., computing Eigenvalues and Eigenvectors and characterizing the corresponding decomposition of a graph's adjacency matrix (or possibly even a tensor of a third dimension is involved [22], e.g., time). For example, PageRank computes the principal Eigenvector of a matrix that is derived from the Web link graph with an additional feature of uniform random jumps. PageRank vectors, characterizing the prestige of Web pages, are usually computed by the iterative Jacobi method (aka. power iteration) with fast convergence. The same mathematical algorithm has been applied to computing TrustRank [12], SpamRank [3], and other methods of social authority and (positive or negative) impact.

This family of spectral analysis procedures includes many

*VLDB '07*, September 23-28, 2007, Vienna, Austria.

variants for different types of graphs, such as directed vs. undirected graphs, weighted vs. unweighted edges, one type of nodes vs. typed or labelled graphs, and so on. One significant drawback of most algorithms of this kind is that they operate on a potentially huge graph and expect the graph to reside in main memory (possibly in the aggregated memory of many computers in a server farm). This may be acceptable for Google, but becomes a major impediment for other applications.

## 1.3 Our System

Many applications of authority analysis naturally lend themselves towards a decentralized peer-to-peer (P2P) setting. For example, tagged photo collections would ideally reside on the owner's computer and shared with the community by a P2P-style network, and the same holds for lists of friends and private interactions, recommendations, and subjective "ratings" within a tightly-knit community. Today's social-Web applications are typically implemented in a centralized server-centric manner for technological as well as business-model reasons, but P2P implementations are not only well conceivable, but would actually have advantages in terms of lower vulnerability to performance bottlenecks, privacy breaches, and other forms of attacks, censorship, or manipulation. These arguments would even hold for Google-style Web search; several research projects are underway to study whether Web search can be implemented in a P2P network with millions of autonomous peers. Such systems could obviously highly benefit from the availability of global authority scores.

We consider a self-organizing network of collaborating peers, using an overlay network. Each peer is autonomous in the sense that it has full control over its own data contents posted to the P2P network. This data may be personal content (such as photos), opinions or recommendations (such as product reviews or blog articles), and other social tags that refer to data residing elsewhere. For example, in a P2P Web search network, a peer's data may be a small collection of, say a few million, Web pages gathered by a focused crawler that is trained with the corresponding user's thematic interest profile. Thus, each peer would have a local fragment of the global social graph that emanates from the overall network. Because of the peers' autonomy, the local graphs of different peers may overlap; so a peer's data may contain Web pages, community members, or other entities (and the corresponding graph vicinities) that are also known to and captured in data of other peers.

This P2P setting complicates PageRank-style authority analysis. Our JXP algorithm [18] is a solution to compute authority measures in a completely decentralized manner. Our fully implemented system provides a versatile platform for different kinds of authority computations in such a P2P environment, and is not limited to any concrete application domain. For very large global graphs that would not fit into the memory of a single computer, JXP has the salient property that it can distribute and parallelize the work across many smaller peers.

# 2. OUTLINE OF THE JXP ALGORITHM

Recently, various techniques have been proposed for distributed PageRank-style authority computations, e.g., [23, 24]. However, these advanced methods work only when the overall Web graph is partitioned into disjoint fragments. With autonomous peers creating (or gathering) and tagging data at their own discretion, this is not a reasonable assumption.

The JXP algorithm dynamically computes global PageRank (or mathematically similar) authority scores based on directed graphs that are arbitrarily spread over autonomous peers in a P2P collaboration. Each peer periodically, and independently of other peers, performs local PageRank score computations on its local graph fragment, where the local graph is augmented by a *world node* that represents the locally unknown part of the global graph. Mathematically, this is a state lumping or aggregation technique for the underlying Markov chain.

JXP initiates random meetings between pairs of peers, for mutual exchange of information about their local graph fragments and to continuously improve each peer's knowledge about its world node. The algorithm does not assume any particular assignment of pages to peers, allowing the local graphs to overlap arbitrarily. Throughout these meetings, the JXP scores that are locally maintained at each peer for its graph fragment converge to the global authority scores that would be derived from the entire global graph.

Details about the JXP algorithm, including the proof of convergence, can be found in [18].

The JXP algorithm is efficient and scalable, as all computations are strictly local and performed only on the small local graph fragments, whose sizes are independent of the number of peers in the network or the size of the global graph.

# 3. EXPLOITING SOCIAL TAGS

Explicitly annotating entities within a community, also referred to as *Social Tagging*, creates peer relationships on different levels:

- The number of annotations of an entity, possibly normalized by its lifetime, indicates the level of interest in the entity. This may be captured in edge weights.
- If the same entity has been tagged by several users, both users seem to share interest in the topic of the entity — not only that they have accessed the entity, but also cared about annotating it.
- If their tags even match, the users do not only seem to share the same interest, but also agree in their opinions.

When meeting previously unknown peers, the level of mutual "social compliance" can be checked by means of statistical synopses and relative entropy measures [19] either on the user's local data or on the similarity of user-assigned tags. A simple thresholding allows the formation of communities of thematically "close" peers, much in the spirit of a semantic overlay network (SON) [7, 1]. Such friend relationships can be used to bias future peer interactions as to personalize the computational results. For example, if you are a computer scientist, you may want to weigh entities created by other computer scientists higher than those from economists or marketing people, or you might bias the weights of an authority analysis in the spirit of a *personal PageRank score*.

The contribution of meeting a remote peer w.r.t. local authority scores positively depends on the number of links to local pages, as these are the paths of authority transfer. To benefit from this observation we bias the meeting strategy of JXP in two ways:

- Since it has been shown that for web documents the majority of outgoing links point to document that are in the same community, we favor peers that have similar content. Preferring such peers when scheduling meetings can, thus, speed up convergence, as they have a higher probability of containing links to pages in the local collection.
- We perform a pre-meeting phase where peers exchange compact data synopses that describe the outgoing links of their collections.

While the first strategy is inherently given in a semantic overlay network, the second strategy is an additional filtering step that prevents peers from performing useless meetings.

## 4. DEMO DESCRIPTION

The demo showcases a collaboration of peers in a self-organizing P2P network. We use Amazon product pages and their recommendations as small-scale data for the live demonstration. We point out that we can apply the very same software also to other social network data, e.g., harvested from Flickr or Del.icio.us, and we are currently in the process of harvesting large-scale experimental data. While this data in principle exhibits a stronger use-case for authority analysis, we see the main purpose of a demonstration at a conference to actually perform a live demonstration of the prototype (instead of only showing screenshots on a poster), we have opted to nevertheless settle for the smaller data set, but hope to make Flickr data available by the time of the conference.

On Flickr data, i.e., assuming annotated pictures residing locally at the peers and an additional user graph built up from contact relations, consider the following use case: when browsing the images, a user comes across an image of the Prater, one of the most renowned sights of Vienna, and is interested in more high-quality pictures of the Prater. While in today's system, a search for the tag *Prater* will probably result in a huge result list with an unclear notion of relevance, the authority scores allow a distinct approach of relevance. Consider the user graph as implicit expressions of user endorsements, just like the document link graph does for Web pages in PageRank. Users with a large number of (incoming) edges in the user graph can be considered to be valuable sources for high-quality pictures, which has triggered their popularity. Authority-scores of the users resulting from a JXP-style distributed authority-computation can now be used to add a well-founded notion of relevance and, thus, help the user to find high-quality images of the prater (in combination with other sources of relevance, e.g., based on text features or image similarity). Alternatively, the existence of the user graph lets the user easily limit her search to her (immediate or $x$-degree) contacts.

Let us alternatively outline a different application of authority on Del.icio.us data. Consider each user is crawling the Web autonomously, starting with its personal bookmarks as crawl seeds. Doing so, each user builds up a local index and is willing to share its index contents to improve the search experiences of remote users. Conceptually, the underlying graph in this case consists of the (hypothetically) combined local Web graphs of all users. The existence of PageRank-style authority scores on the combined Web graph is a big step towards a meaningful document scoring across the autonomous sources. We plan to implement this scenario on top of our P2P Web search prototype [4].

For the actual conference demonstration, this transfers to the following scenario: peers have drawn data from a Web-Service-based extraction of Amazon product pages. The product pages contain user comments as "tags" to the products. Moreover, each product page has a set of links to recommended, similar products. These links form our graph structure. "Ground truth" is given by global authority scores computed upfront on the entire graph

For live demonstration of scale, multiple peers can be run on the same notebook or multiple peers can be started on multiple notebooks, using a standard LAN switch. JXP peers are implemented in Java 5, the peers' data collections (i.e., their local graphs) are stored in a light-weight database system that supports standard SQL. Upon startup, each peer creates a Pastry [21] node and is automatically inserted into the P2P network. Peers implement the *PastryApplication* interface (see http://freepastry.org) and, thus, can be used together with any existing Pastry-based P2P application. This way, the JXP meeting processes can be piggybacked on the application-specific communication. Note that we, in absence of a particular companion P2P application and for the sake of simplicity of running a demo, use Pastry only as a network infrastructure to keep the network connected and to randomly select peers to meet. We could employ any other network infrastructure, and do not rely on any DHT functionality like redundant data storage.

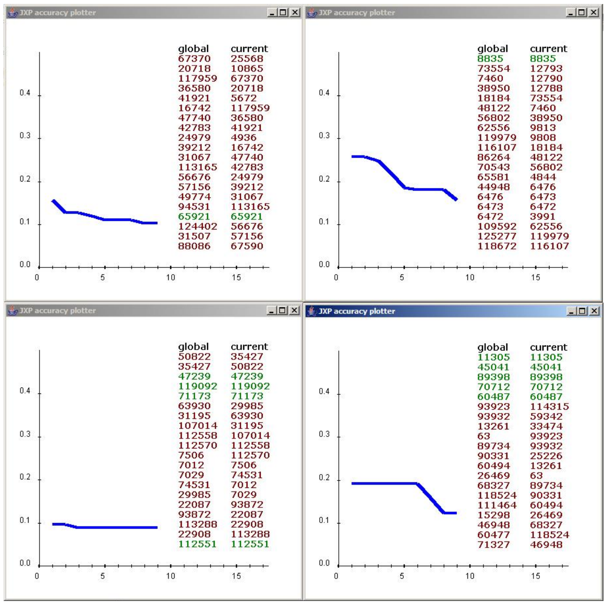
Peers start by computing initial JXP scores on their local graph fragment and gradually improve their knowledge about the outside world by random meetings with other peers. Each peer illustrates its local convergence of JXP scores to the ground truth by charts being updated continuously. Figure 1 shows a screenshot with charts for 4 peers in a network with 10 peers. The y-axis shows Spearman's footrule distance between the ranking of local pages (based on the current JXP scores) and the ground truth, as a function of meetings performed (x-axis). The table on the upper right corner of each graph shows anecdotic evidence in form of the top-20 local pages (again based on the current JXP scores) side-by-side with the top-20 local pages based on the ground truth. Matching ranks are highlighted in green. Notice that the relative ranking of pages is mostly correct even for red pages that are only shifted by one because of a missing or misplaced page above. This gives strong evidence that authority scores, which have been proven to be powerful ingredients to data ranking in a central setting, are finally also available in a fully distributed environment.

Demo visitors can follow the convergence of JXP scores to the ground truth, as the aggregated error is rapidly decreasing and as both rankings align.

The demo can be started in three different modes, i.e. peer meeting strategies, that influence the speed of convergence:

- with remote peers selected randomly and uniformly
- with a bias in favor of peers that have similar content
- using the aforementioned pre-meeting strategy to avoid useless meetings

Demo visitors can easily check that the number of meetings (shown on the x-axis) necessary to reach a satisfactory level of convergence decreases remarkably. This makes a strong case that powerful social systems can also add to the performance of distributed systems.



| global | current |
|---|---|
| 67370 | 25568 |
| 20718 | 10865 |
| 117959 | 67370 |
| 36580 | 20718 |
| 41921 | 5672 |
| 16742 | 117959 |
| 47740 | 36580 |
| 42783 | 41921 |
| 24979 | 4936 |
| 39212 | 16742 |
| 31067 | 47740 |
| 113165 | 42783 |
| 56676 | 24979 |
| 57156 | 39212 |
| 49774 | 31067 |
| 94531 | 113165 |
| 65921 | 65921 |
| 124402 | 56676 |
| 31507 | 57156 |
| 88086 | 67590 |

| global | current |
|---|---|
| 8835 | 8835 |
| 73554 | 12793 |
| 7460 | 12790 |
| 38950 | 12788 |
| 18184 | 73554 |
| 48122 | 7460 |
| 56802 | 38950 |
| 62556 | 9813 |
| 119979 | 9808 |
| 116107 | 18184 |
| 86264 | 48122 |
| 70543 | 56802 |
| 65581 | 4844 |
| 44948 | 6476 |
| 6476 | 6473 |
| 6473 | 6472 |
| 6472 | 3991 |
| 109592 | 62556 |
| 125277 | 119979 |
| 118672 | 116107 |

| global | current |
|---|---|
| 50822 | 35427 |
| 35427 | 50822 |
| 47239 | 47239 |
| 119092 | 119092 |
| 71173 | 71173 |
| 63930 | 29985 |
| 31195 | 63930 |
| 107014 | 31195 |
| 112558 | 107014 |
| 112570 | 112558 |
| 7506 | 112570 |
| 7012 | 7506 |
| 7029 | 74531 |
| 74531 | 7012 |
| 29985 | 7029 |
| 22087 | 93872 |
| 93872 | 22087 |
| 113288 | 22908 |
| 22908 | 113288 |
| 112551 | 112551 |

| global | current |
|---|---|
| 11305 | 11305 |
| 45041 | 45041 |
| 89398 | 89398 |
| 70712 | 70712 |
| 60487 | 60487 |
| 93923 | 114315 |
| 93932 | 59342 |
| 13261 | 33474 |
| 63 | 93923 |
| 89734 | 93932 |
| 90331 | 25226 |
| 60494 | 13261 |
| 26469 | 63 |
| 68327 | 89734 |
| 118524 | 90331 |
| 111464 | 60494 |
| 15298 | 26469 |
| 46948 | 68327 |
| 60477 | 118524 |
| 71327 | 46948 |

**Figure 1: Screenshot for 4 Peers**

## 5. REFERENCES

[1] K. Aberer, A. Datta, and M. Hauswirth. P-grid: Dynamics of self-organizing processes in structured peer-to-peer systems. In *Peer-to-Peer Systems and Applications*, pages 137–153, 2005.

[2] L. Becchetti, C. Castillo, D. Donato, and S. Leonardi. Link-based characterization and detection of web spam. In *AIRWeb*, 2006.

[3] A. Benczur, K. Csalogany, T. Sarlos, and M. Uher. Spamrank - fully automatic link spam detection, 2005.

[4] M. Bender, S. Michel, J. X. Parreira, and T. Crecelius. P2p web search: Make it light, make it fly (demo). In *3rd Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, CA, USA*, pages 164–168, 2007.

[5] K. Berberich, M. Vazirgiannis, and G. Weikum. T-rank: Time-aware authority ranking. In *WAW*, pages 131–142, 2004.

[6] S. Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann, San Francisco, 2002.

[7] A. Crespo and H. Garcia-Molina. Semantic overlay networks for p2p systems. In *AP2PC*, pages 1–13, 2004.

[8] A. Doan, R. Ramakrishnan, F. Chen, P. DeRose, Y. Lee, R. McCann, M. Sayyadian, and W. Shen. Community information management. *IEEE Data Eng. Bull.*, 29(1):64–72, 2006.

[9] D. Gruhl, R. V. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW*, pages 491–501, 2004.

[10] R. V. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW*, pages 403–412, 2004.

[11] Z. Gyöngyi, P. Berkhin, H. Garcia-Molina, and J. Pedersen. Link spam detection based on mass estimation. In *VLDB*, pages 439–450, 2006.

[12] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *VLDB*, pages 576–587, 2004.

[13] H. Hwang, V. Hristidis, and Y. Papakonstantinou. Objectrank: a system for authority-based search on databases. In *SIGMOD Conference*, pages 796–798, 2006.

[14] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *WWW*, pages 640–651, 2003.

[15] V. Krikos, S. Stamou, P. Kokosis, A. Ntoulas, and D. Christodoulakis. Directoryrank: ordering pages in web directories. In *WIDM*, pages 17–22, 2005.

[16] V. Krishnan and R. Raj. Web spam detection with anti-trust rank. In *AIRWeb*, 2006.

[17] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD*, pages 611–617, 2006.

[18] J. X. Parreira, D. Donato, S. Michel, and G. Weikum. Efficient and decentralized pagerank approximation in a peer-to-peer web search network. In *VLDB*, pages 415–426, 2006.

[19] J. X. Parreira, S. Michel, and G. Weikum. p2pDating: Real life inspired semantic overlay networks for web search. Information processing and management (2006), doi:10.1016/j.ipm.2006.09.007.

[20] E. Rahm and A. Thor. Citation analysis of database publications. *SIGMOD Record*, 34(4):48–53, 2005.

[21] A. I. T. Rowstron and P. Druschel. Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. In *Middleware*, pages 329–350, 2001.

[22] J. Sun, D. Tao, and C. Faloutsos. Beyond streams and graphs: dynamic tensor analysis. In *KDD*, pages 374–383, 2006.

[23] Y. Wang and D. J. DeWitt. Computing pagerank in a distributed internet search system. In *VLDB*, 2004.

[24] J. Wu and K. Aberer. Using a layered markov model for distributed web ranking computation. In *ICDCS*, pages 533–542, 2005.