Inferring Emergent Web Communities

Karsten Verbeurgt

Abstract— **One of the characteristic features of the World Wide Web is that it allows anyone to access information on almost any subject that is of interest to that person. The emergent result of this process is that communities arise that are focused around subjects of common interest. Identifying these communities can be challenging, since in contrast to the communities that have been formed more traditionally by people sharing a common interest because they are located in physically adjacent spaces, geographic proximity is not present on the Web. In this paper, we discuss methods of detecting communities on the Web.**

Keywords— **Datamining, Networking**

I. Introduction

THE web is a highly dynamic system, with several million new pages going online every day. Authors of web pages typically link to existing pages on the web that are related to the new page. As a result, the link structure of the web contains information about which pages are related. This leads to the emergence of communities of web pages.

Due to the immense size of the web, mining the link structure of the web to determine the communities is a challenging problem. Several authors have considered graph-theoretic techniques to address this problem [1], [2], [3], [4], [5], [6]. None of these techniques enable a query to determine the community, or communities, that a particular web page belongs to dynamically. The ability to do so is interesting for several applications. For example, this ability would be quite interesting in gathering competitive intelligence regarding the communities that a particular company, perhaps a competitor, is involved in.

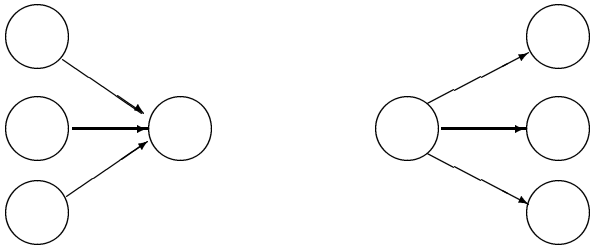
In this paper, we discuss previous notions of communities on the web, and the algorithms that are used to determine those communities. We then propose a new method that, given a web page, will determine the communities to which it belongs dynamically.

II. Finding Communities on the Web

Several different notions of community on the web have been proposed [1], [2], [3], [4], [5], [6]. The common thread amongst the notions of community proposed is that they model the web as a graph based on link structure, and search for properties of the graph deemed to indicate a community of web pages. In this section, we survey the graph properties used to indicate communities, and the algorithms that are used to compute membership in the communities.

Perhaps the most obvious property of a graph that would indicate the presence of a community would be a clique. A clique is a subset of nodes in the graph such that every node is connected to every other node in the subset. There are barriers to finding cliques in the graph to determine communities. The first of these is a computational issue, since the problem of finding the maximum clique in a graph is NP-hard [7]. In addition, it is even hard to find a good approximation to the maximum clique [8]. The second reason against using cliques to determine communities is that the connectivity required in a clique is too strong. It would be too much to expect every member of a community to be linked to every other member of the community.

A notion of community in the web graph that addresses these issues was proposed by Keinberg [3]. Kleinberg's algorithm, called HITS (Hyperlink-Induced Topic Search) views communities as being formed from authorities, to which many pages are linked, and hubs, which link to many pages (Fig. 1). The HITS algorithm computes an "authority weight" and a "hub weight" for each node in the graph.



a) authority node **b) hub node**

Fig. 1. Part a) shows an authority node, with several nodes containing links to the authority. Part b) shows a hub node, that points to many nodes.

The HITS algorithm associates with each page p a *hub weight h(p)* and an *authority weight a(p)*, all initialized to 1. If $p \to q$ represents a link from page p to q, then the weights are adjusted iteratively as follows:

$$a(p) := \sum_{q \to p} h(q)$$

$$h(p) := \sum_{p \to q} a(q)$$

The algorithm first updates the $a(p)$ values based on the $h(q)$ values of pages q pointing to page p, and then updates the $h(p)$ values based on the $a(q)$ values of pages q pointed to by page p.

The HITS algorithm was applied in [2] to extract communities from the web. For this application, HITS was used as a post-processing phase applied to a set of root pages. The root set may in practice be obtained by using a search engine on a particular query to find the root set.

K. Verbeurgt is with the State University of New York at New Paltz, E-mail : verbeurg@cs.newpaltz.edu, Web Address : http://www.cs.newpaltz.edu/~verbeurg .

That set is then expanded by including all forward and backward links to and from the root set. In [2] root sets of 200 pages were used, although the algorithm could easily scale to much larger sets. The core of the community is then defined as the ten pages with the highest $h()$ value, and the ten pages with the highest $a()$ value. The authors in [2] note that the value ten is arbitrary.

As noted by Kleinberg in [3], his HITS algorithm essentially computes the largest eigenvectors of matrices M_{hub} and M_{auth} that are constructed from the connectivity matrix of the web graph. The most authoritative nodes then correspond to the largest entries in the principal eigenvector. It is also noted in [2] that the non-principal eigenvectors can be used to discover additional communities within the set. The principal community corresponds to the set of hubs and authorities that are most densely connected, and the non-principal communities correspond to less densely connected communities.

One characteristic of the above method for determining communities on the web based on the HITS algorithm and its variations is that they consider only link structure. A very interesting challenge is to combine link structure and page content to determine communities. This issue is addressed in [1], where the HITS algorithm is extended to use an affinity measure between pages as a weight on the page link. The affinity value is relative to the fixed query that produced the root set of pages. The affinity value is influenced by the distance of a term of the query from the hyperlink, with closer proximity contributing higher affinity weight. The HITS algorithm is then used on the affinity matrix to compute the principal eigenvector, which is then used to extract the principal community. This is an interesting modification of the HITS technique; however it is limited to constructing communities relative to a query.

All of the variations of the HITS algorithm discussed thus far have been used to extract communities from a relatively small root set that is typically produced via a query on a search engine. Kumar et. al. [4] give an algorithm for "trawling the web" to find communities on the entire web graph. The notion of community that they use is motivated by the hubs and authorities of Kleinberg. In the bibliometric literature, one common measure of how strongly two references are related is how often they occur as co-citations. That is, how often the two references are cited together. For web pages, the idea of co-citation corresponds to separate pages containing links to the same sites. This notion of co-citation is used to define community in [4]: "Web communities are characterized by dense directed bipartite subgraphs." In the terminology of the HITS algorithm, this means that a community is characterized by a set of hubs that all point to the *same* set of authorities (Fig. 2). Thus, this notion of community is actually stronger than the previously discussed notion of hub-authority communities in that it requires all the hubs to point to the *same* authorities.

In [4], the authors show that every large random directed bipartite graph will, with high probability, contain a complete directed (i,j) bipartite graph. (Such a graph contains two sets of nodes: a set of i nodes, each of which has an edge to every node in the set of j nodes.) The authors present a linear-time algorithm to search for complete directed (i,j) bipartite sub-graphs for small values of i and j. These sub-graphs are deemed to indicate the core of a community.

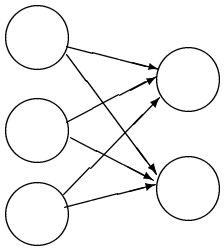


Fig. 2. The core of the community is formed by a set of hubs all connected to the same set of authorities. This forms a (3,2)-bipartite graph as the core of the community.

While the algorithm of Kumar et. al. [4] discussed in the previous paragraphs scales the hub and authority notion of community to large graphs such as the entire web graph, it does have a limitation. Once a node is detected to be a part of a community, it is removed from further consideration. This implies that a page can only belong to one community, which is a considerable limitation of this method.

An additional limitation of all of these notions of community is that they represent a very authoritarian view of the world. If communities are required to contain clear authorities and hubs, then more "democratic" types of communities would be ruled out. For example, a set of nodes in which each node was linked to a somewhat random subset of the nodes in the set would not be considered as a community (Fig. 3). We refer to this type of structure as being "more democratic" because every member has a roughly equal authority measure and hub measure, and there is no consensus among the hubs as to the best authority.

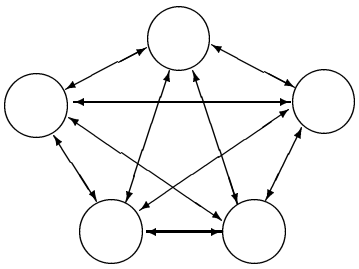


Fig. 3. In structures where each node is connected to a number of other nodes, there may be no clear notion of hubs and authorities. The graph in this figure is an example of a clique structure, where every node is connected to every other node in the set.

That the HITS algorithm does not apply well to this more democratic notion of community is supported by the work of Ng, Zheng and Jordan [9] on the stability of the HITS algorithm. They show that the eigenvectors produced by the HITS algorithm are stable under minor perturbations to the link structure only if there is a relatively large eigen-gap (i.e., a gap between the values in the eigen-

vector.) Since community is being determined by the entries in the eigenvector with the largest value, this implies that the community is stable only when such a large eigengap exists; that is to say, if there are clear indisputable authorities and hubs in the community. Under a more democratic notion of community, this would not be the case, potentially leading to an unstable notion of community. Different authorities and hubs for the community would typically be found on different runs of the algorithm in such cases.

Most of the results in the literature on detecting communities in the web graph are related to the HITS algorithm, as discussed above. We now consider two other results that use different measures of community. The first that we consider is a method by Pirolli, Pitkow and Rao [5]. They use a "spreading activation" function that bears some resemblance to a neural network. A set of "important nodes" are chosen to serve as sources of activation. The activation is then allowed to spread through the graph, so that nodes that are related to the source nodes will receive a higher activation. They incorporate a combination of link structure, textual similarity, and usage statistics in the weights of links in the graph. While this is a potentially interesting notion of determining community, it suffers in that it is not clear that the spreading activation function will "converge" to a set of nodes. From the authors' description, it appears that whether the activation dies out or spreads to the entire graph is highly sensitive to the parameters of the model.

Another interesting technique was proposed in [6]. Given a graph on a set of nodes, the task of dividing the set into sub-graphs, corresponding to sub-communities, is considered. The method proposed used a max-flow algorithm to find the minimum cut in the graph. The minimum cut represents the smallest set of articulation points in the graph that when removed will partition the graph into two disjoint sub-graphs. This method differs from all of the others that we have discussed in that its goal is not to determine a community as a subset of a potentially large graph. The goal of this method is, given a set of nodes, to partition the set into sub-communities. Nevertheless, this method is potentially interesting as a post-processing phase of a community-finding algorithm in order to refine the community into sub-communities.

III. The Connectivity Structure of the Web

Now that we have discussed notions of community on the web that occur in the literature, along with the corresponding algorithms for computing the members of the community, we now turn to a discussion of the connectivity patterns on the web. The connectivity on the web graph plays an important role in determining what types of graph properties can be used in practice to define communities on the web. For example, one might consider proposing connected components of the web graph as a natural property to determine communities. That is to say, find a set of pages that are all connected to each other via forward links. As it turns out, however, this notion of community would in many cases generate communities of several million web pages. This notion of connectivity does identify a particular type of community, which is basically the well-connected web community comprising a large portion of the web. In this section, we further discuss such results on connectivity of the web.

In one of the earlier papers on web connectivity [10], several types of connectivity are discussed. We refer to the web graph as W, and the undirected version of this graph as W_u. W_u is formed from W in a natural way by replacing all edges with an undirected edge. The results of [10] state that W_u contains a "giant component", although they do not quantify how large that component is. They further report that under strong connectivity in W, the largest strongly connected components are small, of size less than 20. More recent results, presented in [11], show the situation to be much more intricate. Their studies show that in fact there is a single strongly connected component at the core of the web that contains about 56 million nodes (Fig. 4). In contrast, the next largest strongly connected components are small in comparison, at about one hundred thousand nodes. In addition, there is a large set of approximately 44 million nodes that have paths leading into the strongly connected central core component, and another set of roughly the same size whose nodes are pointed to by the strongly connected central core. These results were computed on a snapshot of the web that contained 200 million pages. Thus, about 25% of the pages on the web are found to be in the strongly connected central core component. The implication of these results to determination of communities is that under the property of strongly connected components, a large proportion of the web belongs to this single central community at the web's core. Thus, strong connectivity is too "weak" for determining communities at a more local level.

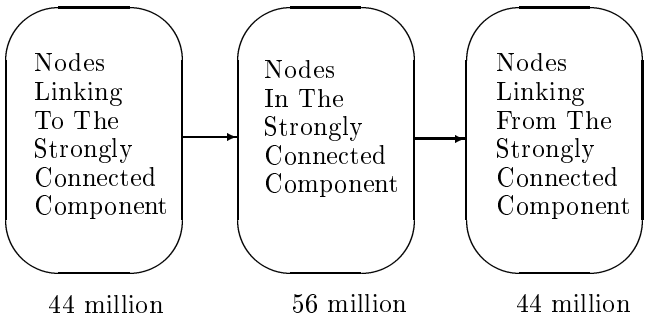


Fig. 4. The strongly connected component that forms the core of the web has approximately 56 million nodes. Thus, there is a path from each of the nodes in this set to every other node in the set. There is also a set approximately 44 million nodes that are not strongly connected, but that contain a path of links that lead into the strongly connected core set. Similarly, there is a set of approximately 44 million nodes that are not strongly connected, but that can be reached from nodes in the strongly connected core. This result shows that the notion of strong connectivity induces immense connected components.

Two other notions of connectivity are discussed in [10]: biconnectivity and alternating connectivity. Biconnectivity applies to the undirected graph W_u, and is defined as

follows: two nodes x and y are *biconnected* if there is no third node z such that z is on every path between x and y. Under this notion of connectivity, the authors find that the web graph contains a "giant biconnected component". They also find that this biconnected component generally contains all of the top hubs an authorities computed by the HITS algorithm.

Since the relation of biconnectivity induces a huge graph, a more concise notion of connectivity might refine the size of the community. Thus, the notion of alternating connectivity was proposed in [10]. An *alternating path* from nodes x to y is a path where the directions of edges strictly alternate between forward an backward (Fig. 5). The *alternating distance* between nodes x and y is defined to be the length of the shortest alternating path between u and v. The *undirected distance* is the length of the shortest path between x and y in the undirected graph W_u. The authors report that the alternating distance in the giant connected component of W_u is generally at most twice the undirected distance. This tends to indicate that there are many alternating paths in W. The significance of this result is that the notion of alternating paths is related to the concept of hubs and authorities. An alternating path allows a zig-zag between hubs an authorities, traversing first from a hub to an authority, and then following a back-link to another hub (Fig. 5). The evidence that there are many alternating paths supports the notion of the web being organized as hubs and authorities. As mentioned in the previous section, however, the hubs and authorities structures do not accommodate some notions of community on the web that seem natural and interesting.

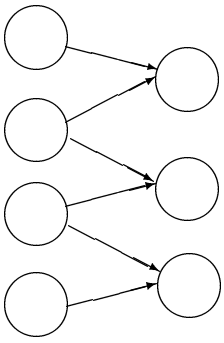


Fig. 5. An alternating path consists of a strict alternation between forward edges and backward edges. Note the similarity between alternating paths and biconnected components, as depicted in Fig. 2

IV. Aspects of Web Community

We have now reviewed several works on determining communities on the web, and have also discussed several results on the connectivity structure of the web. Predominant in these works is the notion of hubs and authorities as forming the core of the community. This notion of community is very interesting, for the following reasons. First, there is an algorithm (namely HITS [3]) to compute the hubs and authorities of the community locally from a root set obtained via a web search. The algorithm is known to converge rapidly in practice. Secondly, there is an algorithm to find communities at a global level ([4]) using this notion of community. Finally, this notion of community is plausible sociologically, in that web content creators do try to include links to authoritative sites, hence creating hubs.

Despite these salient features, the authority-hub notion of community is not all-encompassing. It does not allow for more democratic community structures, as discussed previously in this paper. In addition, if the algorithm is applied to graphs that do not contain the authority-hub structure, the hubs and authorities chosen to form the community will be unstable, as discussed previously and in [9]. Thus, exploring alternate definitions of community to complement the hubs-authorities view of community is an important area for further research. In the remainder of this paper, we discuss criteria for definitions of community, and propose notion of community based upon cyclical structure of sub-graphs.

Before discussing criteria for a definition of community on the web, we first briefly discuss the obvious graph properties to look for that allow more democratic community structures. The first of these is the clique, which is a subset of nodes in the graph such that each node is connected to every other node. There are two problems with using cliques. First, determining the maximum clique is NP-hard [7], and it is hard even to approximate [8]. Secondly, clique connectivity is too much to expect in a web community. Every node would be required to connect to every other node in the community.

The weaker notions of connectivity commonly used in graph theory, such as strongly connected components and weakly connected components, cannot be used in practice as definitions of community. As discussed in the previous section, the results of [11] show that a large fraction of the web is a single strongly connected component of tens of millions of nodes.

Thus, we desire a notion of local connectivity that is somewhere between clique and strongly connected component. It should embody the notion that members of the community are somewhat densely interconnected.

V. Criteria for Definition of Web Community

We are thus in search local properties of the web graph that characterize community, and that have associated algorithms to compute the members of the community. Such a property and its associated algorithm should have the following properties:

1. **Convergence:** The search for the nodes in the community should converge to a set of nodes that meets a definition of a local graph property. As noted previously, strong connectivity would not achieve this, since the computation of the community would not halt in many cases until tens of millions of nodes were included in the community.
2. **Efficiency:** The algorithm for searching for a community must be efficient. For example, maximum clique would not meet this criteria, since there are no efficient algorithms known for maximum clique.

3. **Naturalness:** The notion of community should bear some relation to a natural notion of community. For example, dense connectivity amongst the nodes in the community would be a property that could be interpreted as being naturally indicative of the presence of a community.

It is worth noting that the hubs and authorities definition of communities conforms to the three properties mentioned. In fact, these salient properties of definitions of community were motivated by their presence in the work on communities of hubs and authorities.

VI. A New Definition of Community

We seek to propose a definition of community that will allow for more democratic community structures than the hubs and authorities model of communities, as discussed previously in this paper. This notion of community will conform to the intuition that a community is a somewhat densely connected set of nodes, and will also conform to the three properties outlined above, that we argue are essential for any meaningful definition of community.

The motivation for our definition of community is the intuition that if a node is a member of a community, then it will be linked to the rest of the community, and it will also have links from the rest of the community. Thus, if a web user follows a link to another member of the community, they would eventually find their way back to the initial node if it is indeed a community. The relevant graph-theoretic structure to this intuition is a cycle in a graph. A *cycle* in a graph starting at node x is a directed path that begins at x and ends at x, for which all intermediate nodes are distinct. The length of the cycle is the number of links in the cycle. To detect a community involving a node x, we first choose a maximum diameter of the community we are interested in finding. In rare cases this may be the diameter of the entire web, but in general, it will be much smaller. We then apply a variant of depth first search called depth-limited search, which is commonly used in the artificial intelligence literature [12]. This algorithm conducts a depth-first search up to a fixed depth. The depth we will use is the maximum community diameter discussed above.

Using the depth-limited search, we propose to find all cycles of length bounded by the community diameter using the standard depth-first search algorithm for determining cycles in a graph [13]. When a cycle is found, each node along the cycle is credited with a weight inversely proportional to the length of the cycle in which it occurs. That is to say, if a cycle is found of length t, then each node along the cycle is credited with a weight of $1/t$. This heuristic is designed to give higher credit for smaller cycles, implying tighter communities. After the depth-limited search has been completed to the diameter of the community, we are guaranteed that all cycles of length less than or equal to the specified diameter involving node x will have been found, and the weight at node x will indicate the strength of its membership in the community. The value accumulated at node x is a measure of how well it is connected to the other nodes searched via cycles. This value can then be thresholded to determine communities of a specified strength. This allows us to model a continuum of increasing levels of connectivity, from strongly connected components to clique connectivity.

The method of determining community that we propose here meets all three of the criteria we outline for a definition of community. Due to the bounded diameter of the search, the method is guaranteed to halt at the specified diameter, hence meeting the convergence criteria. It is also efficient both in time and memory usage, since it conducts a depth-limited search. Finally, it corresponds to a natural notion of community; that of node x being connected to the rest of the community by a path that leads back to the node through the community.

This new notion of community is presented here as a preliminary work. We have not at this time implemented the method and run experiments using it to validate this notion of community. It is our intention to do so in future work.

VII. Conclusion

In this paper, we have surveyed results on determining communities in the web. Most of the literature pertinent to this area stems from Kleinberg's work on hubs and authorities as the core of web communities. This is a very interesting notion of community on the web, and experimental work validates that indeed much of the web is structured as hubs and authorities. However, the hubs and authorities notion of communities does not accommodate some graph structures, which we refer to as "more democratic" community structures where all members of the community have a more equal status rather than centering around well-defined authorities. These types of community structures are interesting for several applications. For example, it could be used to discover communities in hyperlinked newsgroup discussions. Another interesting application is competitive analysis in e-business. Given the web site of particular company, finding the web communities of which it is a member may reveal interesting information about the alliances of that company with others on the web.

We intend to further investigate local graph properties that lead to the determination of communities in the web in future research.

References

[1] Soumen Chakrabarti, Byron Dom, David Gibson, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins, "Spectral filtering for resource discovery," in *ACM SIGIR workshop on Hypertext Information Retrieval on the Web*, 1998.

[2] David Gibson, Jon M. Kleinberg, and Prabhakar Raghavan, "Inferring Web Communities from Link Topology," in *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, Pittsburgh, Pennsylvania, June 1998, pp. 225–234.

[3] Jon M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[4] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins, "Trawling the Web for emerging cyber-communities," *Computer Networks (Amsterdam, Netherlands: 1999)*, vol. 31, no. 11–16, pp. 1481–1493, 1999.

[5] Peter Pirolli, James Pitkow, and Ramana Rao, "Silk from a sow's ear: Extracting usable structures from the web," in *Proc.*

ACM Conf. Human Factors in Computing Systems, CHI. 1996, ACM Press.

[6] Gary William Flake, Steve Lawrence, C. Lee Giles, and Frans M. Coetzee, "Self-organization and identification of web communities," *IEEE Computer*, pp. 66–71, March 2002.

[7] M.R. Garey and D.S. Johnson, *Computers and Intractability*, W.H. Freeman and Company, 1979.

[8] Johan Hastad, "Clique is hard to approximate within n," *Acta Mathematica*, vol. 182, no. 1, pp. 105–142, 1999.

[9] A. Ng, A. Zheng, and M. Jordan, "Link analysis, eigenvectors and stability," in *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, Seattle, Washington, August 2001, pp. 903–910.

[10] Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S. Tomkins, "The Web as a graph: Measurements, models and methods," *Lecture Notes in Computer Science*, vol. 1627, pp. 1–18, 1999.

[11] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Weiner, "Graph structure in the web," http://www9.org/w9cdrom/160/160.html.

[12] Stuart J. Russel and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 1995.

[13] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest, *Introduction to Algorithms*, The MIT Press, 1990.

Karsten Verbeurgt obtained his PhD in Computer Science from the University of Waterloo, Canada, in 1998. Since that time, he has been an Assistant Professor at the State University of New York at New Paltz. Dr. Verbeurgt's research interests include information retrieval, search engine design, data mining, machine learning, and artificial intelligence. Dr. Verbeurgt can be reached via E-mail at verbeurg@cs.newpaltz.edu, or via his web page at http://www.cs.newpaltz.edu/~verbeurg .