Web Social Mining

Hady W. Lauw, Ee-Peng Lim

Keywords: Web social mining, Web 2.0, social network discovery, social network analysis, social network applications.

Abstract

With increasing user presence in the Web and Web 2.0, Web social mining becomes an important and challenging task that finds a wide range of new applications relevant to e-commerce and social software. In this article, we describe three Web social mining topics, namely, social network discovery, social network analysis and social network applications. The essential concepts, models and techniques of these Web social mining topics will be surveyed so as to establish the basic foundation for developing novel applications and for conducting research.

1 Introduction

Web social mining refers to conducting social network mining on Web data. Here, we adopt a very broad interpretation of Web data which includes Web sites, Web pages, Web servers' and applications' log data, as well as user-generated data from Web 2.0[1] sites. As increasing amount of user data is made available on the Web, it opens up a new world of opportunities for the Web data to be mined for realizing new applications and making existing ones work more intelligently.

As shown in Figure 1, web social mining can be covered in three aspects, namely, *social network discovery*, *social network analysis*, and *social network applications*. Social network discovery refers to the construction of social networks linking users and sometimes other semantic entities together so as to study individual- or community-level properties in social network analysis. Patterns and knowledge about individuals and their communities are then incorporated into a wide range of social network applications.

While web social mining poses more diverse opportunities for commercial applications, it has a deep root in social network analysis, a research discipline pioneered by social scientists. Hence, many of the models and techniques developed for social network analysis by social scientist are still applicable to web social mining. On the other hand, web social mining has added new challenges of automatically discovering social networks from the raw web data which we call social network discovery.

The objective of this article is to survey the essential concepts, problems, solution techniques and applications of web social mining. Hopefully, this will serve as a good introduction to web social mining and a reference for future research and application development. In this article, we give an overview of web social mining by first examining the various forms of Web data

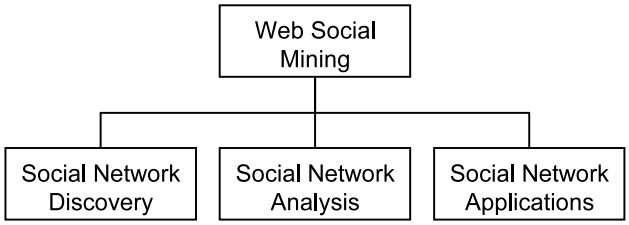

Figure 1: Web Social Mining Topics

available for social network mining. We then introduce a set of fundamental social network concepts. We review the web social mining work in three subsequent sections, covering social network discovery, analysis, and application respectively. Given that web social mining covers a large set of concepts and topics, we shall only describe the key ones very briefly. Interested readers can refer to the provided references for more detailed information.

2 Web Data Sources

Web social mining can be conducted on a plethora of web data embedding information about user-user and user-object links. Traditional web data sources consists of web pages from different sites, as well as the user browsing and search activity records logged by web servers, web applications (e.g., web search engines, e-commerce sites, etc.) and web browsers. Web page data are often regarded as unstructured content documents in which people, company, product and other entity names may be found and their relationships can be extracted by text mining. In some websites, web pages may be much more structured as the pages are directly generated from data maintained in relational or XML databases. An example of such websites is the DBLP Computer Science Bibliography[1] (or simply DBLP). DBLP provides bibliographic information of computer science publications organized by author, conference, journal and subject. When websites contain structured content about semantic entities, their data can potentially be used for web social mining. In the case of DBLP, there have been much social network mining research on co-authorships among researchers since one can easily extract the co-authors of publications[2, 3].

Web social mining actually begins to flourish when Web 2.0[1] becomes popular. Web 2.0 consists of Internet sites that offer web users a range of services to interact with one another, sharing information, collaborating, and maintaining social relationships. As Web 2.0 sites attract huge population of users, there are also commercial incentives drawing upon the social relationships among users to further enhance user experiences at these sites, and/or to generate revenues from advertisement or product sales. This can be done by discovering the influence of users' opinions, providing new services to users (e.g., product recommendation), etc.. In the following, we classify the existing Web 2.0 sites into four broad categories by the characteristics of their data.

[1] http://www.informatik.uni-trier.de/~ley/db/

- Social networking sites: Examples of social network sites include Facebook[2], MySpace[3] and Linkedin[4]. These are Web 2.0 sites allowing users to construct their personal profiles as well as to connect themselves with networks of friends. As the relationship links among users at these sites are user specified, they usually provide the ready social networks for further analysis. One can also correlate the network properties (e.g., authority) with the personal profile attributes.

- Content sharing sites: Web 2.0 sites for content sharing include YouTube[5], Flickr[6], delicious[7], and many others. The content to be shared cover video, audio, photo images, social bookmarks, etc.. Using these sites, users publish their content files making them easily accessed, commented and rated by other users. These content sharing sites offer large set of content objects in addition to user data for constructing large social networks and determining the user interests and other properties in the networks.

- Collaboration sites: There are several Web 2.0 sites offering collaboration services to users. Here we highlight two typical collaboration examples, namely Wikipedia and community question answering (QA) portals, e.g., Yahoo! Answers[8], askville[9] and answerbag[10]. Wikipedia is currently the largest online encyclopedia with millions of articles collaboratively edited by millions of users. In community QA portals, users post questions and other users answer them. As multiple answers can be provided to the same questions, one can find collective efforts in answer contribution. At the collaboration sites, each user leaves a trace of his or her contribution (e.g., authored article content, questions, answers) which can be used for web social mining.

- E-Commerce sites: E-commerce sites such as eBay[11], yelp[12], and Epinions.com[13] are beginning to harness user participation to create new business models that create new revenues. For example, eBay relies on buyers rating sellers so as derive the latter's reputation. Epinions and yelp, on the other hand, have users providing reviews and ratings on products. While E-commerce sites have tighter control over their data, they often provide rating and pricing information about products which can be used in web social mining.

3 Fundamental Concepts

We review the basic terminology of social network that will be used for the rest of the article.

[2]http://www.facebook.com
[3]http://www.myspace.com
[4]http://www.linkedin.com
[5]http://www.youtube.com
[6]http://www.flickr.com
[7]http://delicious.com
[8]http://answers.yahoo.com
[9]http://askville.amazon.com
[10]http://www.answerbag.com
[11]http://www.ebay.com
[12]http://www.yelp.com
[13]http://www.epinions.com

Actor

An actor is an entity whose relationships to other actors are mapped onto a social network. Examples of actors include people, objects, organizations, countries, etc.

Link

A link directly relates a pair of actors. There could be diverse meanings attached to a link, including: *evaluation* (e.g., liking/disliking, respect, friendship), *affiliation* (e.g., person belonging to a club), *interaction* (e.g., communicating, collaborating), etc.

A link is either *directed* from one actor to another, or *undirected* if it is symmetrically shared between the two actors. A *dichotomous* link is either present or absent, while a *valued* link is weighted with a range of values, with higher values usually indicating stronger relationships. A valued link may also be *unsigned*, with positive link weights, or *signed*, where link weight may be positive or negative (e.g., liking or disliking).

Path

A path connects a pair of actors through an unbroken chain of links. The length of a path is the number of links that make up the chain.

Subgroup

A subgroup comprises a subset of actors in a social network, as well as all the links between them. The actors to be included in a subgroup are selected based on specific criteria, which will be discussed later.

Relation

A social network may have several types of links. A relation is the set of all links of a specific type. For example, if we define two relations $\mathcal{R}_{friend}$ and $\mathcal{R}_{work}$, then all links based on friendship make up $\mathcal{R}_{friend}$ and all links based on working relationship make up $\mathcal{R}_{work}$.

Mode

A social network may have several types of actors. Mode refers to the number of distinct types of actors. If all actors are of the same type (e.g., people), the network is a one-mode network. If there are two types of actors (e.g., people and organizations), it is a two-mode network.

4 Social Network Discovery

The problem of social network discovery can be expressed as follows: *given a finite set of actors, find out which pairs of actors have a link between them and, if applicable, what the weight of each link is.* The solution to this problem requires some criterion to decide whether there is sufficient evidence to infer a link between two nodes and to quantify the strength of that link. Below, we list four such criteria that have been used in prior work, namely: *self-reported*, *communication*, *similarity*, and *co-occurrence*. As shown by the taxonomy in Figure 2, the former two usually

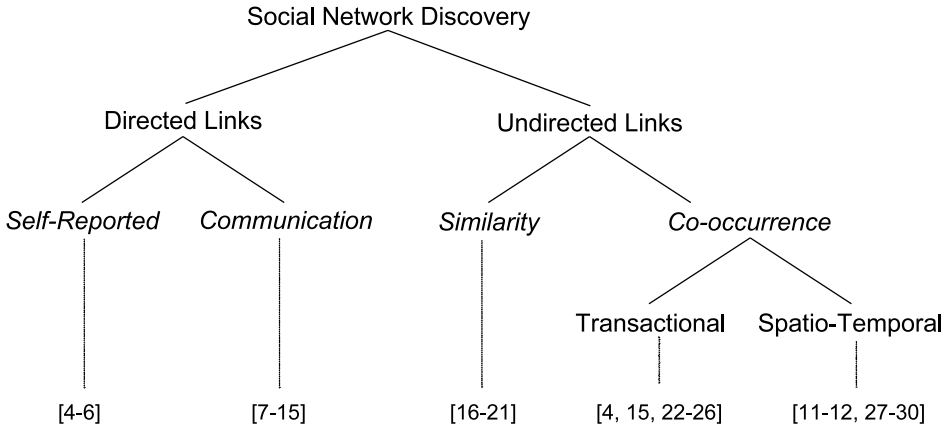

Figure 2: Taxonomy of Social Network Discovery

give rise to directed links; the latter two, to undirected links. Note that for each criterion, links can be inferred from either offline or online activities.

4.1 Self-Reported

Self-reported links refer to links discovered from the involved actors themselves. A directed link from actor a_i to another actor a_j exists if a_i has reported it. Such links are directed since a_j may not necessarily report a link to a_i. Even if a pair of actors mutually report links to each other, they may not attach equal weights to the link.

Classical social network research discovers self-reported links through carefully constructed procedures such as questionnaires, interviews, direct observations of interactions, manual sifting through archival record, or various experiments [4]. The discovery effort is time- and resource-intensive, covers a small number of actors, and is usually restricted to specific settings (e.g., people in a company/school).

Web settings lower the barrier and create incentives for a user to report links to others. Someone maintaining a homepage or a blog often lists hyperlinks to Web sites or blogs of friends (e.g., LiveJournal [5]), to increase her connectivity within the community, which helps to increase traffic to their homepage or blog. Similarly, profile pages of community-centric sites such as Facebook or Friendster [6] commonly display a self-professed list of friends within the community. Consequently, there are voluminous and diverse self-reported links that can be harvested from these sources.

4.2 Communication

Communication, defined generally as transfer of information or resources, is commonly exhibited by socially related people. Communication-based links are usually directed from the originator to the recipient. If desired, an undirected link may be inferred from bi-directional links. Links are usually weighted by the frequency and intensity (e.g, conversation length) of the communication.

Evidence of communication can be drawn from direct observation of interactions or interviews, e.g., asking a group of people to give accounts of work communication [7]. Much of modern communication is computer-mediated, over the Internet, which often leaves a trail in the form of usage logs that can be mined for evidence of sustained communication. Sources of online communication include records of email [8, 9], Instant Messaging (IM) [10, 11, 12], newsgroups [13, 14], phone logs [15], etc.

4.3 Similarity

Similarity has its foundation on the well-received sociological idea that friends tend to be alike [16, 17]. This leads to the premise that the more people have in common, the likelier it is that they are related. Similarity-based links are naturally undirected, since the notion of similarity is symmetric.

Prior work on similarity-based links involves identifying the relevant attributes of users that may indicate relationship, and a suitable similarity measure. Homepages with similar content and linkages may represent a group of related individuals [18]. Two people whose sets of communication partners overlap may be affiliated to a common group [19]. Other forms of similarity include sharing the same opinions or areas of interest [20], or even sharing similar vocabulary choices in email messages [21].

4.4 Co-occurrence

Co-occurrence assumes that if several actors occur together more frequently than random chance alone would allow, they are likely associated in some way. Like similarity, it is also undirected by nature. Prior work on co-occurrence-based links can be organized into two streams: *transactional*, where there is a clear boundary within which two actors are said to co-occur, and *spatio-temporal*, where the boundary of co-occurrence is defined by space and/or time.

Transactional Co-occurrence

The term *transaction* is borrowed from work on frequent pattern mining [22, 23]. It refers to a discrete instance within which a few items may co-occur, e.g., a supermarket transaction involving a number of product items. A frequent pattern involves a set of items that co-occur together in many transactions, and thus are likely to be associated with one another. Applied to social network discovery, a transaction in an offline setting may refer to a party attended by a pair of actors [4], a movie that a pair of actors act in [15], or a publication which a pair of researchers co-author [24, 25]. In an online setting, a transaction may refer to a Web page where the names of a pair actors co-occur [26].

Spatio-Temporal Co-occurrence

The boundary of a transaction is not always clear-cut, especially when it involves continuous dimensions such as space and time. Suppose that we have a set of tuples $\{\langle a, s, t\rangle\}$, where each tuple records an actor a appearing at location s at time t, and we wish to infer links between pairs of actors based on co-occurrences. A transaction must then be defined in terms of space and/or time. For example, a spatial transaction can be derived by discretizing the

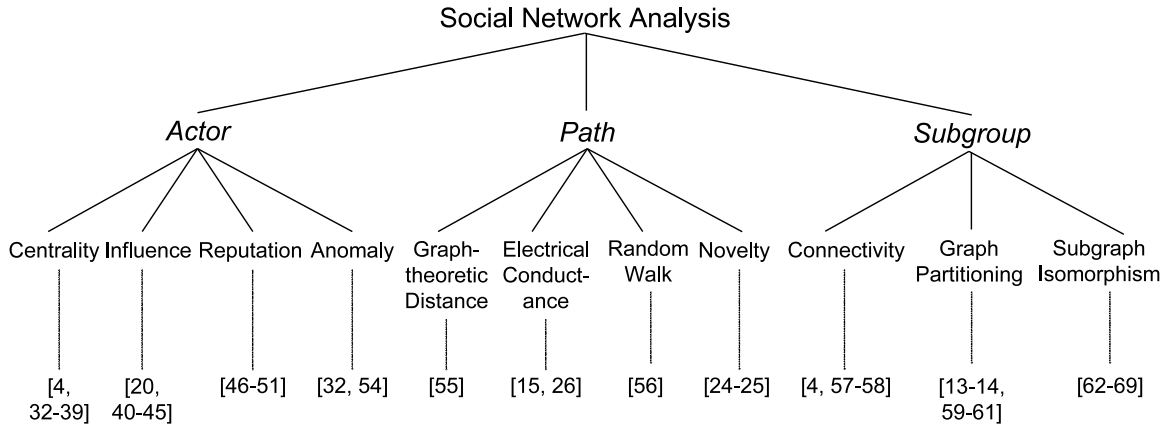

Figure 3: Taxonomy of Social Network Analysis

space dimension using a sliding window [27]. A temporal transaction can be a time interval within which two IM users must be online together (and thus are more likely to engage in a conversation) [11, 12].

In turn, a spatio-temporal co-occurrence is defined over both space and time. That spatio-temporal movement data is a possible indicator of social association has been suggested in [28, 29, 30]. Our work *STEvent* in [31] concerns social network discovery from spatio-temporal co-occurrences. STEvent focuses on the analysis of movement data and algorithm development to infer associations. It generalizes the spatio-temporal co-occurrence beyond movement over physical locations to include other location types such as cyber locations.

5 Social Network Analysis

Social network analysis attempts to find useful structures, patterns, or insights that exist within a social network. As shown in the taxonomy in Figure 3, such studies may look for "important" actors in the network (*actor analysis*), "important" paths connecting a subset of actors (*path analysis*), and subgroups that exist within a network (*subgroup analysis*).

Note that we do not distinguish between social networks derived from offline or online activities. Most analytical methods simply assume a readily available social network. Neither do we distinguish between directed links or undirected links. Most analytical methods can be adapted to both types of links. The common workaround is to define analysis for directed links and treat undirected links as bi-directional links, or to define analysis for undirected links and ignore the direction of directed links.

5.1 Actor Analysis

The problem of actor analysis can generally be expressed as follows: *given a social network, measure or rank the "importance" of every actor in the network.* There are various definitions of importance, which usually represents a certain property or behavior of an actor. As shown

in the taxonomy in Figure 3, prior work in actor analysis has largely focused on the following definitions of importance: *centrality*, *influence*, *reputation*, and *anomaly*.

Centrality

Centrality equates importance of actors to occupying strategic or central locations in a network [4]. Such actors are more visible and are involved in more relationships with other actors. Social network researchers have developed the following measures of centrality, that are mostly based on the structural properties of a graph.

Degree. The degree centrality of an actor is her number of links. The intuition is that central actors should be the most active, and should have the most connections to others in its vicinity. This measure has been applied to law enforcement, where it is used to identify the key players in a price fixing conspiracy [32], and the supposed ringleader of 911 terrorist network (Mohammed Atta) [33].

Closeness. The closeness centrality of an actor is the inverse of the average path length from the actor to all other actors in the network. The reasoning is that an important actor should have easy access to others members of the network.

Betweenness. The betweenness centrality of an actor is the number of distinct shortest paths (connecting any pair of actors) that pass through it. Actors with high betweenness values are in a position to control communication channels, either by impeding or accelerating or just by getting informed of such communication.

Eigenvector Centrality. The eigenvector centrality of an actor is the sum of the eigenvector centralities of other actors with links to the actor [34, 35]. This measure takes into account not just the number of links that an actor has, but also the quality of those links. Intuitively, a central actors is one whom many other central actors link to. The most well-known and successful application of eigenvector centrality is for ranking Web pages based on hyperlinks for Web search, e.g, PageRank [36], HITS [37], and various other link analysis algorithms [38, 39].

Influence

Influence equates importance of actors to ability to propagate the adoption of an idea or a product to other actors in the network. The mode of propagation could be through various channels such as word-of-mouth or persuasion. This measure founds application in viral marketing, which depends on identifying high-influence individual to promote products and services to their acquaintances [40, 41, 42, 43, 20].

The propagation framework is as follows [41]. Each actor is in one of two states: active or inactive. Initially, only one or a few *seed* actors are active, while the rest are inactive. The propagation of active state proceeds in discrete iterations. In each iteration, an inactive actor may get activated by its active neighbors. Actors that are active in the previous iterations remain active. The iterations terminate after a preset number of iterations, or when no further activation is possible. The influence of an actor (or a small subset of actors) is measured by using the actor(s) as seed actor(s) and counting the final number of active actors at the end of the iterations. The mechanism by which an actor is activated generally falls into either the *threshold model* or the *cascade model*.

Threshold Model. In the threshold model [44], each actor a_j has a threshold activation value of θ_j, and the link weight w_{ij} from a_i to a_j reflects a_i's degree of influence on a_j. Actor a_j is

activated in the iteration when $(\sum_{\text{active } a_i \in neighbors(a_j)} w_{ij}) \geq \theta_j$.

Cascade Model. In the cascade model [45], the link weight w_{ij} from a_i to a_j reflects the probability that a_i can successfully activate a_j. In each iteration of the propagation process, each active actor a_i is given a chance to activate an inactive neighbor a_j with a probability of success equal to w_{ij}.

Reputation

Reputation is often equated with trustworthiness. In online settings, interaction between strangers is common. Thus, platforms that support such interactions (e.g., online auction sites) often institute a reputation system that allows users to evaluate how trusted an actor is by others in the network. All things being equal, one would rather transact with actors of higher reputation. There are two main criteria for inferring the reputation of an actor: *past behaviors* and *trust evaluation by others*.

Past Behaviors. One way to determine how trustworthy an actor will be in the future is to see how trustworthy the actor has been in the past. The auction site eBay maintains a *feedback score* for each registered user. On completing a transaction, a buyer and a seller may give a feedback point to each other, which can be 1 (positive rating), 0 (neutral rating), or -1 (negative rating). The feedback score (reputation) of an actor is his/her running total of feedback points [46]. In product review site Epinions, a user may write product reviews and get paid based on the number of people who read the reviews. Each review may also be rated by other users. The reputation of a user is a function of the rating scores received by the user's reviews [47].

Trust Evaluation by Others. Some systems such as FOAF [48] and Epinions [49] maintain a social network, where each link denotes a trust relationship. Thus, another way to determine how trustworthy an actor is is to see how many other actors in the network trust her [50, 51]. For example, the work on EigenTrust [50] measures the reputation of an actor as the sum of the reputations of other actors with trust links to the actor (akin to eigenvector centrality applied on a network of trust relationship).

Anomaly

In contrast to centrality, anomaly equates importance to being different from or having few connections to other actors. For instance, key players (bosses) in a criminal network may intentionally keep a distance from others for fear of detection by the police and let their underlings carry out their wishes [32]. Finding anomalous actors is akin to outlier detection [52, 53], which is concerned with identifying data points that are situated at a distance from the majority of data points. In prior work, anomalous actors have been defined as those with low closeness centrality values [32], or those least visited by random walks starting from other actors in the network [54].

5.2 Path Analysis

The problem of path analysis can generally be expressed as follows: *given a social network and* ≥ 2 *seed actors, identify the set of "important" paths connecting the seed actors.* The important paths are those that are most likely undertaken from one seed actor to another. Prior work is organized based on how each defines what make up the important paths. As shown

in the taxonomy in Figure 3, the four main criteria are: *graph-theoretic distance*, *electrical conductance*, *random walk*, and *novelty*.

Graph-theoretic Distance

Several distance measures in graph theory that could serve to measure the importance of a path include *shortest path* and *maximum flow*.

Shortest Path. The shortest path is the path with minimum number of links (for dichotomous links), or the path with maximum total weight (for valued links). This measure has been used to identify strongest association paths between entities in a criminal network [55]. For instance, if two criminals are known to be cooperating, they are likely to use the shortest path between them. Individuals along this association path are themselves potential suspects in the criminal activity.

Maximum Flow. In the maximum flow approach, the social network is modeled as a flow graph. One seed actor is designated the source node, and the other the sink node. Each link in the network is a channel for the flow of material, which is limited by the capacity (link weight). The maximum flow path allows the greatest flow of materials from the source to the sink.

Electrical Conductance

A social network could also be modeled as an electrical circuit. Each seed actor is assigned a potential (source node 1V and sink node 0V). Each link is like a resistor with a certain conductance value (link weight). The best path is the one that delivers the highest electrical current from the source node to the target node. The electrical conductance model for mining interesting connections between individuals in a social network was first proposed by [26], and further improved upon by [15].

Electrical conductance is superior to graph-theoretic distance measures in two ways. Unlike the shortest path approach, this model takes into account the popularity of intermediate nodes in a path. Popular nodes allow greater leakage of electricity, corresponding to weaker and incidental connections to a popular person that a normal person would have. Unlike the maximum flow approach, this model takes into account the length of a path in determining interestingness. Longer paths accumulate resistance which impedes the flow of electricity, similar to weaker social relationship to be expected from a longer social path.

Random Walk

Another way to measure path importance is using the random walk mechanism. Random walk is a traversal of a social network graph, which starts from a seed actor and picks the next neighboring actor to visit randomly (either with uniform probability or with probability proportional to link weight). If we start independent random walks from each seed actor, intuitively the paths that are most commonly traversed by these random walks in aggregate are the most important paths connecting the seed actors. The work on center-piece subgraph [56] applies the random walk model to find interesting co-authorship connections. Unlike the electrical conductance model, the center-piece subgraph may also include good paths that connect only a subset of (not all) seed actors.

Novelty

Path importance may also be defined in terms of novelty or uniqueness. A given social network may consist of a few relations (e.g., friendship relation, work relation). Thus, a path may be constructed by links of a few different relations. The novelty of a path is how rarely the combination of relation types in its links can be found in other paths. A novel path captures a unique and exclusive relationship between the seed actors. For example, [24] discovered paths denoting student-teacher relationships based on their exclusive co-authorship with each other. [25] found that the only two mafia groups to be involved in a gang war in a simulated criminal database were connected by paths made up of novel combinations of evidence links (e.g., money transactions, meetings, killings).

5.3 Subgroup Analysis

In a social network, for every actor, there is a relatively small subset of other actors that the actor knows well; that small subset constitutes a subgroup. In general, members of a subgroup interact more frequently and intensively with other members than with non-members. A network consists of one or more subgroups, which may or may not overlap with each other. The subgroup analysis problem can be concisely stated as follows: *given a social network, identify the subgroups in the network.* In prior work, there are various definitions of what constitutes a subgroup. As shown in the taxonomy in Figure 3, these definitions fall into one of three categories: *connectivity*, *graph partitioning*, and *subgraph isomorphism*.

Connectivity-based Subgroups

Connectivity-based subgroups are defined in terms of how connected members in a subgroup are [57, 58, 4]. Here we look at three such criteria: mutuality, reachability, and nodal degree.

Mutuality. Mutuality-based subgroups, called *cliques*, are maximal complete subgraphs of at least three actors. This definition captures the idea of cohesiveness, where everyone knows everyone else. However, due to its strictness, cliques are relatively rare in real-life data.

Reachability. Reachability only requires that any pairwise members of a subgroup is reachable from each other through a path of a length not more than n links. If the path may involve an actor outside the subgroup, the subgroup is called *n-clique*. A more restrictive version, *n-clan*, can be derived by rejecting those n-cliques that require a path involving a non-member.

Nodal Degree. Another way to relax the mutuality requirement is to allow each actor to have a lower degree than mutuality would have required. Given k and n, a subgroup of n members is termed a *k-plex* if at most k links can be missing from each actor to its neighbors, or a *k-core*, if at least k links must be present from each actor to its neighbors.

Graph Partitioning

Graph partitioning assumes that a social network consists of a set of disjoint subgroups. Finding those subgroups involves removing a set of links such that the social network graph is partitioned into disjoint subgraphs. This method has been used to find subgroups in networks with unsigned links as well as those with signed links.

Unsigned Links. In a network of unsigned links, the objective is to partition the graph into components, such that each component is relatively dense, but the cut (the set of links to be

removed) between any two components is relatively sparse. As there could be many possible cuts, the best cut is the one that maximizes the value of some goodness function. This method has been used to partition a collection of newsgroups [14] and Web pages [59, 60] into subgroups consisting of newsgroups or Web pages of similar topics.

Signed Links. In a network of signed links, the objective is to partition the graph into components, by removing negative links, such that each component consists of as many positive links as possible. For example, [13] split contributors of newsgroups on controversial issues (e.g., politics, abortion) into two camps: those who are for or against a particular issue. [61] split a network of political parties and a network of tribes into subgroups of similarly aligned parties/tribes.

Subgraph Isomorphism

Subgraph isomorphism assumes that a subgroup has a non-random pattern of linking among its members (subgraph pattern), which is shared by a number of other subgroups. Hence, finding subgroups within a network is equivalent to finding subgraph patterns that have many isomophic instances in the network. Below, we review two approaches to derive such subgroups: *Apriori-like algorithms* and *compression-based approach.*

Apriori-like Algorithms. A subgraph pattern is frequent if the number of isomophic instances meets the specified threshold value. To reduce the space of subgraph patterns whose frequencies have to be determined, most of the proposed algorithms [62, 63, 64, 65, 66, 67] follow the general principle of the Apriori algorithm that was first proposed by [23] for mining association rules from transaction databases. Adapted to graph data, the principle states that a subgraph pattern has a higher frequency than any of its supergraphs (other patterns that subsume the subgraph). If a subgraph pattern is not frequent, none of its supergraphs need to be considered.

Compression-based Approach. Unlike the apriori-like algorithms that find all subgraph patterns whose frequencies meet the threshold, the compression-based approach employs a greedy algorithm to find a subset of subgraph patterns that together result in a good compression of the original graph [68]. Using the Minimum Description Length (MDL) principle, compression is achieved by replacing all isomorphic instances of a subgraph pattern with a more concise representation called "concept". [69] used this approach to identify substructures in a terrorist network, revealing the chain-like communication channels used by terrorist cells.

6 Social Network Application

Below, we list a number of applications (mostly online applications) with web social mining aspects. While the list is by no means exhaustive, it sufficiently paints a picture of how the techniques reviewed earlier in this article may be used in real-life applications.

Online Social Media

Online social media refers to online applications for disseminating and sharing information that also support socially-oriented features. Examples of such applications include: blogs (e.g., LiveJournal[14]), wikis (e.g., Wikipedia), content sharing (e.g., Flickr for photos, YouTube for videos),

[14] www.livejournal.com

online communities (Facebook [70], Friendster [6], MySpace [71]), and social bookmarking (e.g, delicious). Such applications often allow users to assign *tags* (textual annotations) to objects in order to collaboratively organize content, to assign *ratings* to collaboratively evaluate content, and to maintain one's *social network* in order to track the latest goings-on, activities, and interests of friends. The social aspects of these activities lend themselves to social network analysis. For example, by analyzing the pattern of hyperlinking among blog posts, we can identify the opinion leaders among bloggers [72]. By analyzing the edit history of Wikipedia articles, we can identify the most authoritative authors [73].

Social Search

Social search refers to querying one's social network to look up interesting actors or paths. For instance, one may look for actors whose profile fit the description given in a query, e.g., someone looking for potential dates [70]. Alternatively, one may look for actors holding a specific piece of information [74, 75]. This is especially useful for information that is not widely available and may not be indexed in public databases. For example, the answer to the question "Which camera shop in my local neighborhood would offer a good deal to students of my university?" is probably known by a university friend who is an avid photographer. One may also search for interesting association paths. *ReferralWeb* [76] allows a user to explore the chains of referrals leading to a target actor. Users of such a system may be a businessman who wishes to get an introduction to a potential business partner or a graduating student who needs a reference letter from a well-known academician.

Recommender Systems

Recommender systems are online applications that generate personalized recommendations (e.g., which book to buy) based on information provided by the users [77, 78, 79]. Some recommender systems require the user to manually enter a personal profile of interests, preferences, or expertise. Others may infer this information implicitly from the user's past activities, e.g., user's purchasing history at Amazon[15] or user's ratings on movies at GroupLens[16]. A similarity-based social network can then be constructed based on this information. The system could then generate recommendations to an actor based on what other similar or related actors have purchased or rated highly.

Academic Peer Review

Peer review refers to the collaborative exercise in which academicians evaluate each other's work, in order to determine which papers should be accepted for publications in conference proceedings and journals, or which reseach proposals should be granted funding. Questions that often come up during the peer review process include how to identify the best papers or proposals taking into account the varying rating scores assigned by different reviewers [80, 81], and how to best assign reviewers to objects (papers or proposals) taking into account such factors as the match in topics between reviewers and objects and the workload of reviewers [82, 83, 84].

[15] www.amazon.com
[16] www.grouplens.org

Social network techniques would likely be useful in deriving the answers to these questions as many academic activities can be mapped onto social network representation. For example, there is a wealth of reseach on social networks based on co-authorship [85, 15, 24, 56], co-citation (being cited together in publications) [86, 87, 88, 89, 90, 91], bibliographic coupling (citing common publications) [92], etc. Social network analysis can be employed to generate insights that would help to improve and inform the peer review process, e.g., identifying the authorities in specific fields [91], or tracking which communities are growing or shrinking [85].

7 Conclusions

Web social mining is a topic that sees the cross-fertilization of computing and social science leading to a wide range of interesting applications on the Web. This article provides a brief survey of the essential concepts and techniques used in Web social mining. It covers social network discovery that allows social networks to be derived from Web and Web 2.0 data, social network analysis that find patterns and knowledge about actors, paths and other structures in the social networks, and some example applications that can benefit from Web social mining. As new forms of Web data and applications emerges, new Web social mining models and techniques will be in demand thus inspiring more vibrant research in this area.

References

[1] T. O'Reilly. What is web 2.0: Design patterns and business models for the next generation of software. *O'Reilly*, 2005.

[2] Osmar R. Zaiane, Jiyang Chen, and Randy Goebel. Dbconnect: mining research community on dblp data. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 74–81, 2007.

[3] M.A. Nascimento, J. Sander, and J. Pound. Analysis of sigmod's co-authorship graph. *SIGMOD Record*, 32(3):8–10, 2003.

[4] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[5] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogspace. *Communications of the ACM*, 47(12):35–39, 2004.

[6] D. M. Boyd. Friendster and publicly articulated social networking. In *Extended abstracts of the Conference on Human Factors and Computing Systems*, pages 1279–1282, Vienna, Austria, 2004.

[7] L. Garton, C. Haythornthwaite, and B. Wellman. Studying online social networks. *Journal of Computer-Mediated Communication*, 3(1), June 1997.

[8] A. Chapanond, M. S. Krishnamoorthy, and B. Yener. Graph theoretic and spectral analysis of Enron email data. In *Proceedings of the Workshop on Link Analysis, Counterterrorism, and Security (in conj. with SIAM International Conference on Data Mining)*, pages 15–22, Newport Beach, CA, USA, 2005.

[9] J. Diesner and K. M. Carley. Exploration of communication networks from the Enron email corpus. In *Proceedings of the Workshop on Link Analysis, Counterterrorism, and Security (in conj. with SIAM International Conference on Data Mining)*, pages 3–4, Newport Beach, CA, USA, 2005.

[10] M. A. Ahmad and A. Teredesai. Modeling spread of ideas in online social networks. In *Proceedings of the 5th Australasian Conference on Data mining and Analystics*, pages 185–190, Darlinghurst, Australia, 2006. Australian Computer Society, Inc.

[11] J. Resig, S. Dawara, C. M. Homan, and A. Teredesai. Extracting social networks from instant messaging populations. In *Workshop on Link Analysis and Group Detection (in conj. with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining)*, Seattle, WA, USA, 2004.

[12] J. Resig and A. Teredesai. A framework for mining instant messaging services. In *Workshop on Link Analysis, Counterterrorism, and Privacy (in conj. with SIAM International Conference on Data Mining)*, Lake Buena Vista, FA, USA, 2004.

[13] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th International World Wide Web Conference*, pages 688–703, Budapest, Hungary, 2003.

[14] C. Borgs, J. Chayes, M. Mahdian, and A. Saberi. Exploring the community structure of newsgroups. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 783–787, New York, NY, USA, 2004. ACM.

[15] Y. Koren, S. C. North, and C. Volinsky. Measuring and extracting proximity in networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 245–255, New York, NY, USA, 2006. ACM.

[16] S. L. Feld. The focused organization of social ties. *American Journal of Sociology*, 86:1015–1035, 1981.

[17] K. Carley. A theory of group stability. *American Sociological Review*, 56(3):331–354, 1991.

[18] L. A. Adamic and E. Adar. Friends and neighbors on the Web. *Social Networks*, 25(3):211–230, July 2003.

[19] M. F. Schwartz and D. C. M. Wood. Discovering shared interests using graph analysis. *Communications of the ACM*, 36(8):78–89, 1993.

[20] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 61–70, Edmonton, Alberta, Canada, 2002.

[21] P. S. Keila and D. B. Skillicorn. Structure in the Enron email dataset. In *Proceedings of the Workshop on Link Analysis, Counterterrorism, and Security (in conj. with SIAM International Conference on Data Mining)*, pages 55–64, Newport Beach, CA, USA, 2005.

[22] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM International Conference on Management of Data*, pages 207–216, Washington, USA, 1993.

[23] R. Agrawal and R. Srikant. Fast algorithm for mining association rules. In *Proceedings of the 20th International Conference on Very Large Databases*, pages 487–499, Santiago, Chile, 1994.

[24] S. Lin and H. Chalupsky. Unsupervised link discovery in multi-relational data via rarity analysis. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, pages 171–178, Melbourne, FL, USA, 2003.

[25] S. Lin and H. Chalupsky. Issues of verification for unsupervised discovery systems. In *Workshop on Link Analysis and Group Detection (in conj. with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining)*, Seattle, WA, USA, 2004.

[26] C. Faloutsos, K. S. McCurley, and A. Tomkins. Fast discovery of connection subgraphs. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 118–127, Seattle, WA, USA, 2004.

[27] S. Shekhar and Y. Huang. Discovering spatial co-location patterns: A summary of results. In *Proceedings of the 7th International Symposium on Spatial and Temporal Databases*, pages 236–256, Redondo Beach, CA, USA, 2001.

[28] T. Choudhury and A. Pentland. Sensing and modeling human networks using the Sociometer. In *Proceedings of the 7th IEEE International Symposium on Wearable Computing*, pages 216–222, Washington, DC, 2003. IEEE Computer Society.

[29] N. Eagle and A. Pentland. Reality Mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

[30] M. Terry, E.D. Mynatt, K. Ryall, and D. Leigh. Social Net: Using patterns of physical proximity over time to infer shared interests. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems*, pages 816–817, New York, 2002. ACM.

[31] H.W. Lauw, E.-P. Lim, T.T. Tan, and H. Pang. Mining social network from spatio-temporal events. In *Workshop on Link Analysis, Counterterrorism and Security at SDM'05*, 2005.

[32] J. Xu and H. Chen. Untangling criminal networks: A case study. In *Proceedings of the 1st Symposium on Intelligence and Security Informatics*, pages 232–248, Tucson, AZ, USA, 2003.

[33] V. E. Krebs. Mapping networks of terrorist cells. *Connections*, 24(3):43–52, 2002.

[34] P. Bonacich. Technique for analyzing overlapping memberships. *Sociological Methodology*, 4:176–185, 1972.

[35] P. Bonacich. Simultaneous group and individual centralities. *Social Networks*, 13(2):155–168, 1991.

[36] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. In *Stanford Digital Library Technologies Project*, 1998.

[37] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[38] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Link analysis ranking: Algorithms, theory, and experiments. *ACM Transactions on Internet Technology*, 5(1):231–297, 2005.

[39] T. H. Haveliwala. Topic-sensitive PageRank: A context-sensitive ranking algorithm for Web search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784–796, 2003.

[40] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 57–66, San Francisco, CA, USA, 2001.

[41] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, Washington, D.C, USA, 2003.

[42] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1):5, 2007.

[43] K. Ong, W. Ng, and E. Lim. Mining relationship graphs for effective business objectives. In *Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 561–565, Taipei, Taiwan, 2002.

[44] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1987.

[45] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, 2001.

[46] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman. Reputation systems. *Communications of the ACM*, 43(12):45–48, 2000.

[47] M. Chen and J. P. Singh. Computing and using reputations for internet ratings. In *Proceedings of the 3rd ACM Conference on Electronic Commerce*, pages 154–162, New York, 2001. ACM.

[48] J. Golbeck and J. Hendler. Inferring binary trust relationships in Web-based social networks. *ACM Transactions on Internet Technology*, 6(4):497–529, 2006.

[49] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th International Conference on World Wide Web*, pages 403–412, New York, NY, USA, 2004. ACM.

[50] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The Eigentrust algorithm for reputation management in P2P networks. In *Proceedings of the 12th International Conference on World Wide Web*, pages 640–651, New York, NY, USA, 2003. ACM.

[51] L. Xiong and L. Liu. Peertrust: Supporting reputation-based trust in peer-to-peer communities. *IEEE Transactions on Knowledge and Data Engineering*, 16(7):843–857, 2004.

[52] A. Arning, R. Agrawal, and P. Raghavan. A linear method for deviation detection in large databases. In *Proceedings of the 2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 164–169, New York, 1996. ACM.

[53] E. Knorr and R. Ng. A unified notion of outliers: Properties and computation. In *Proceedings of the 3rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 219–222, 1997.

[54] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *ICDM*, pages 418–425, 2005.

[55] J. Xu and H. Chen. Fighting organized crimes: Using shortest-path algorithms to identify associations in criminal networks. *Decision Support Systems*, 38(3):473–487, 2004.

[56] H. Tong and C. Faloutsos. Center-piece subgraphs: Problem definition and fast solutions. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 404–413, New York, NY, USA, 2006. ACM.

[57] L. C. Freeman. The sociological concept of "group": An empirical test of two models. *American Journal of Sociology*, 98:152–166, 1992.

[58] L. C. Freeman. Cliques, galois lattices, and the structure of human social groups. *Social Networks*, 18(3):173–187, 1996.

[59] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, New York, NY, USA, 2000. ACM.

[60] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee. Self-organization and identification of Web communities. *Computer*, 35(3):66–71, 2002.

[61] B. Yang, W. Cheung, and J. Liu. Community mining from signed social networks. *IEEE Transactions on Knowledge and Data Engineering*, 19(10):1333–1348, 2007.

[62] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Computing Surveys*, 38(1):2, 2006.

[63] A. Inokuchi, T. Washio, and H. Motoda. An apriori-based algorithm for mining frequent substructures from graph data. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 13–23, London, UK, 2000. Springer-Verlag.

[64] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 313–320, Washington, DC, USA, 2001. IEEE Computer Society.

[65] M. Kuramochi and G. Karypis. Discovering frequent geometric subgraphs. *Information Systems*, 32(8):1101–1120, 2007.

[66] Xifeng Yan and Jiawei Han. gSpan: Graph-based substructure pattern mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, page 721, Washington, DC, USA, 2002. IEEE Computer Society.

[67] X. Yan and J. Han. CloseGraph: Mining closed frequent graph patterns. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 286–295, 2003.

[68] D. J. Cook and L. B. Holder. Graph-based data mining. *IEEE Intelligent Systems*, 15(2):32–41, March 2000.

[69] M. Mukherjee and L. B. Holder. Graph-based data mining on social networks. In *Workshop on Link Analysis and Group Detection (in conj. with the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining)*, Seattle, WA, USA, 2004.

[70] C. Lampe, N. Ellison, and C. Steinfield. A Face(book) in the crowd: Social searching vs. social browsing. In *Proceedings of the 20th ACM Conference on Computer Supported Cooperative Work*, pages 167–170, New York, 2006. ACM.

[71] S. Patil and J. Lai. Who gets to know what when: Configuring privacy permissions in an awareness application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 101–110, New York, NY, USA, 2005. ACM.

[72] X. Song, Y. Chi, K. Hino, and B. Tseng. Identifying opinion leaders in the blogosphere. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management*, pages 971–974, New York, NY, USA, 2007. ACM.

[73] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong. Measuring article quality in Wikipedia: Models and evaluation. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management*, pages 243–252, New York, NY, USA, 2007. ACM.

[74] B. Yu and M. P. Singh. Searching social networks. In *Proceedings of the 2nd Joint Conference on Autonomous Agents and Multiagent Systems*, pages 65–72, New York, 2003. ACM.

[75] J. Zhang and M. van Alstyne. SWIM: Fostering social network based information search. In *Extended Abstracts on Human Factors in Computing Systems*, pages 1568–1568, New York, 2004. ACM.

[76] H. Kautz, B. Selman, and M. Shah. Referralweb: Combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65, 1997.

[77] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[78] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53, 2004.

[79] P. Resnick and H. R. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[80] Tracy Riggs and Robert Wilensky. An algorithm for automated rating of reviewers. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 381–387, New York, 2001. ACM.

[81] Henry M. Walker, Weichao Ma, and Dorene Mboya. Variability of referees' ratings of conference papers. In *Proceedings of the 7th Annual Conference on Innovation and Technology in Computer Science Education*, pages 178–182, New York, 2002. ACM.

[82] S. Dumais and J. Nielsen. Automating the assignment of submitted manuscripts to reviewers. In *Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 233–244, 1992.

[83] J. Geller and R. Scherl. Challenge: Technology for automated reviewer selection. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 55–61, 1997.

[84] S. Hettich and M. J. Pazzani. Mining for proposal reviewers: Lessons learned at the National Science Foundation. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 862–871, 2006.

[85] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 44–54, New York, NY, USA, 2006. ACM.

[86] C. Chen and L. Carr. Trailblazing the literature of hypertext: Author co-citation analysis (1989-1998). In *Proceedings of the 10th ACM Conference on Hypertext and Hypermedia*, pages 51–60, New York, NY, USA, 1999. ACM.

[87] M.J. Culnan. The intellectual development of management information systems, 1972-1982: A co-citation analysis. *Management Science*, 32(2):156–172, 1986.

[88] D. Sullivan, D.H. White, and E.J. Barboni. Co-citation analyses of science: An evaluation. *Social Studies of Science*, 7:223–240, 1997.

[89] H. Small. Co-citation in the scientic literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, 1973.

[90] H. G. Small. A co-citation model of a scientific specialty: A longitudinal study of collagen research. *Social Studies of Science*, 7:139–166, 1977.

[91] H.D. White and K.W. McCain. Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of American Society of Information Science and Technology*, 49(4):327–355, 1998.

[92] M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963.