An Overview of Social Network Extraction Systems

2.1 Introduction

Social media has changed the way user driven content and information is produced, transmitted and consumed [Leskovec, 2014]. In an online social media setup, content is generated when someone performs a status update, posts a blog, comments on a post or status, tweets or retweets, etc. This content acts as a link between these social actors i.e. producers and consumers of information [Leskovec, 2014]. This link or relationship is the most important piece of information for social network analysis. Social Network Analysis has a lot of potential for businesses and other consumers of information. Tracking and analyzing the flow of information on online social media provides a means for companies to gain feedback on their products or services. This information helps them to improve their products and services, market those products in a better way, and maintain competitive advantage [Gundecha and Liu, 2012; Leskovec, 2014]. Participants on these networks have the advantage of making much more informed decisions by using the wisdom of crowds on these networks [Leskovec, 2014]. This can be attributed to the availability of vast and diverse amount of information on these networks.

There are a number of techniques in literature that attempt to leverage this link information and infer or extract social networks from the underlying data. These techniques use diverse information sources containing this link
information either explicitly or implicitly. The information sources used by these techniques vary from highly unstructured webpages on the one hand to highly structured e-mail communications on the other [Arif et al., 2014a].

Beside their potential utility for business organizations [Leskovec, 2014], extraction and analysis of social networks can also help answer a number of questions related to academics [Arif et al., 2014a]. In the later part of 1990’s researchers started to infer and extract academic social networks automatically. Starting with efforts made by Henry Katz and company [Katz et al., 1997], improvements have been made in the quality and efficiency of the extracted academic social networks. The extracted social networks have been able to answer a number of questions related to networks and information flow. These questions include: who is the main player in a network (hub or leader); who has more connections; how strong the collaboration ties are; how collaborative the authors are; etc. [Arif et al., 2012a].

Figure 2.1 outlines the broad steps commonly used for social network extraction and analysis.

Figure 2.1: Commonly used approach for social network extraction and analysis [Bjelland, 2014].

2.2 Social Network Extraction

As they say, “a picture is better than a thousand words”. Data expressed through figures, tables and projections is more amenable to understanding than numbers or text. It has been observed that human eyes are better at analysing and extracting patterns from figures, tables and pictures [Bjelland,
One can use Social Network Analysis (SNA) to systematically analyse social networks. SNA has long been used as a means of satisfying our needs of understanding our ties within the community of which we are a part [Bjelland, 2014]. Various social network analysis metrics like centrality, betweenness, clustering coefficient, etc. [Arif et al., 2012a] have been used to answer some of the questions mentioned in Section 2.1. To answer these questions and analyse these networks through SNA we need to first extract them.

This process starts with the extraction of actors and their relationships from the underlying datasets. Data mining techniques can be used to extract either explicit or implicit information from the target dataset [Gundecha and Liu, 2012]. This process is called Social Network Extraction (SNE) [Bjelland, 2014]. The methodology adopted for social network extraction depends upon the type of social network being extracted. From business point of view or for customer analysis one needs customer data readily available in databases [Bjelland, 2014]. But if we need to consider multiple relationships or data has to be obtained from diverse and open sources, we may require querying a search engine or using a crawler to extract the required information. However, one has to be very clever to extract valid data in an efficient manner [Bjelland, 2014]. Mining the required data in such a situation is made difficult by the size and diversity of the Web and issue like name ambiguity and understanding contextual information [Bjelland, 2014].

### 2.2.1 Extracting Academic Social Networks

Extraction and analysis of academic social networks involves three major activities: relationship extraction, name disambiguation and profile integration. Among them name disambiguation is the most difficult and important one. This can be inferred from the statistics shown in Chapter 1. The efficacy of social network extraction system and profile integration is dependent of the efficiency of name disambiguation.
In 1997, a humble attempt was made by the architects of Referral Web [Katz et al., 1997] to extract egocentric networks of academics from various information sources like personal homepages, organizational websites, co-authored publications, citations, etc. using basic similarity measures. Since then, a number of attempts have been made to extract meaningful and informative social networks of interest from various online sources. In this Chapter we briefly discuss the most relevant social network extraction techniques. This is followed by a discussion on name disambiguation on which network extraction depends critically.

2.3 Social Network Extraction Techniques

2.3.1 Information Sources for Extraction of Social Networks

The Web has been expanding with great pace which means more information as well as information sources. It has become gigantic and diverse. The social network extraction techniques discussed in this Section have used the following Web information sources, either individually or in combination. These include:

- Webpages
- E-mail communications/logs
- Instant messaging
- Internet relay chats
- Blogs
- Online social networking sites
- News
- Web albums

2.3.2 Techniques and their Classification

The techniques proposed so far extract social networks from explicit as well as implicit relationship information. The social networks extracted from these diverse information sources include academic social networks,
conference participant networks, networks of actors, e-mail communication networks, etc. The existing network extraction techniques differ from each other in number and type of information sources used, the type of networks they seek to extract, etc. However, to the best of our knowledge, there is no formal classification of the existing extraction techniques. We have stepped in to fill this gap and we propose a classification of these techniques on the basis of the information source they use.

2.3.2.1 Proposed Classification of SNE Techniques

We classify these social network extraction techniques based on the type of the information source used for relationship information as follows:

- Web-based Social Network Extraction
- E-mail-based Social Network Extraction
- Instant Messaging/Chat-based Social Network Extraction
- Blog-based Social Network Extraction
- Online Social Networking Sites -based Social Network Extraction
- Multisource data-based Social Network Extraction

Next, we briefly discuss various representative social network extraction techniques on the basis of this classification.

2.3.3 Representative Techniques

2.3.3.1 Web-based Social Network Extraction

Majority of the social network extraction methods [Tombe et al., 2003; Katz et al., 1997; Matuso et al., 2006; Mika, 2005] which use Web as information source, use the co-occurrence of two named entities on a Web source. The frequency of such co-occurrence is indicative of the strength of the relationship. The strength of this relationship is measured from search results returned by a search engine for a query containing these two names.
Chapter 2: An Overview of Social Network Extraction Systems

Academic social network extraction started with the web-based SNE technique Referral Web [Katz et al., 1997] which was extended by [Tombe et al., 2003; Mika, 2005; Matuso et al., 2006]. The co-occurrence measure is the fundamental idea behind [Katz et al., 1997; Tombe et al., 2003; Mika, 2005; Matuso et al., 2006]. The co-occurrence measure, relation identification, and threshold, are determined in advance. This approach works well for homogeneous entities however it may not perform well for heterogeneous entities [Tombe et al., 2003]. Co-occurrence of the target entities on a large number of Web sources also serve as a bottleneck for techniques based on such as measure [Tombe et al., 2003].

The ever increasing data on the Web brought into focus the need to distinguish similar entities in such a way that only information about the relevant entity is extracted to derive a social network out of it. For this, the social network extraction method must be able to address the problems of profile extraction which in turn depends primarily upon name disambiguation mechanism employed. Arnetminer8 [Tang et al., 2007; Tang et al., 2008] tried to address these issues for academic researchers by using Conditional Random Fields (CRFs) [Lafferty et al., 2001] to extract researcher profiles and associated publications data from the Web. Tang et al., [2007] found name disambiguation as the most important and hard to address problem. This has been the focus of Arnetminer also, which points to its significance. This project has tried to address it in a series of studies [Zhang et al., 2007; Wang et al., 2008; Wang et al., 2010b; Wang et al., 2011; Tang et al., 2012].

Poliquen and Atkinson [2008] extract social networks from news articles by using co-occurrence of named entities as a link between them. Though, most of the techniques discussed in this Section used co-occurrence of entities as a measure to compute the weight of the extracted relations, some

---

8 Arnetminer: www.arnetminer.org
of them have examined how to weigh each relation among them [Matuso et al., 2006]. Oka and Matuso [Oka and Matsuo, 2009] propose a method for weighting the relation among entities which assigns different weights to different relations, thus overcoming some of the shortcomings of the co-occurrence based metrics.

2.3.3.2 E-mail Communication-based Social Network Extraction

Email is one of the primary ways that people use to communicate. Because of its inherent properties, it is considered as a promising area of work on communities and social networks [Tyler et al., 2003]. It is the number one online activity for most users [Cullota et al., 2004]. Various studies [Tyler et al., 2003; Van and Zhang, 2003; Cullota et al., 2004; Bird et al., 2006] use e-mail communications to extract social networks from this highly structured information source. Ubiquity of email usage; frequency, longevity, and reciprocity of email communications; type (content) of communication; temporal data; and availability on both sender and receiver side make it a rich source for extraction of communication data. In some cases [Tyler et al., 2003], only header data is used to extract a social network, whereas in some cases [Van and Zhang, 2003], contents in message body have also been used. Techniques like [Van and Zhang, 2003] compromise the privacy and confidentiality of the message, which may be of some concern. These concerns can be addressed by accessing header information only. The information contained in the message body can be ignored to address privacy and confidentiality issues. However, ignoring information contained in message body significantly limits the potential of using email as source of information for analyzing social relationship. Mixed approaches like [Cullota et al., 2004] obtain basic information from e-mail message headers and contact and other related information from the Web.

In spite of being a highly structured information source, the potential of e-mail communications as a data source for social network extraction gets
limited because of certain issues like: multiple identities of same person; spam and group aliases; categorization of social relations by content in message body; using reciprocity, frequency of communication, and longevity of discussion for weighting this relationship; and aggregation of “To, Cc, Bcc” [Van and Zhang, 2003]. The technique proposed in [Bird et al., 2006] tries to overcome some of these issues, like resolution of aliases, by employing a hybrid (manual & automatic) name disambiguation mechanism. Of late it has been observed that social media is gaining more prominence as a means of instant communication. E-mail now a days is used as a means of communication only when formal communication is required.

2.3.3.3 Instant Messaging/Chat-based Social Network Extraction

Instant messaging (IM) or Internet Relay Chat has been a popular form of real time computer-based communications service. Relationship identification and extraction is a central problem in the analysis of such large-scale social networks as there is no clear measure of relationship strength. In [Mutton, 2004] several such measures, obtained from the status log of an IM user, have been proposed that describe the link information between any pair of members. Relationship identification from status logs is a difficult task notwithstanding its apparent simplicity [Resig et al., 2004]. The difficulty can be alleviated by obtaining acquaintances (e.g. buddies in AOL) list for each user but unfortunately, such lists are not public. List owners need to be contacted to obtain these lists which seem impractical. The solution lies in constantly tracking the status (online, busy, away, offline etc.) of each user relative to the IM service. It is possible to track status and transition (user state transitions) from electronically published IM data. In [Mutton, 2004], these status logs are used to measure the degree of relationship between any two AOL IM\(^9\)users.

\(^9\)www.aim.com
In [Resig et al., 2004], an Internet Relay Chat (IRC) bot called PieSpy\(^\text{10}\) is used to monitor IM channels to infer the social network structure. Measures like, direct addressing of users, temporal proximity, temporal density, and private message monitoring have been used to infer relationship strength. After inferring social relations, it [Resig et al., 2004] uses modified spring embedder force model based on [Fruchterman and Reinglod, 1991] for connected network components and m-limited force model [Resig et al., 2004] for disconnected components networks, for extracting social networks.

2.3.3.4 Blogs-based Social Network Extraction

Blogosphere is a virtual collection of social media sites which provide a platform for individuals to express their ideas, discuss various events, share their opinions, facts, events, etc. These discussions range from personal life to society, politics to religion, science and technology to superstition, etc. This environment stands out as a virtual network and is considered a rich source of social information. The dramatic increase in their size, diversity and popularity in recent years has made it a ripe field for automatic extraction of underlying social networks. Contextual similarity between two named entities in sentences of a blog is used as a measure of their being related in [Mesquita et al., 2010] and entities sharing a certain predefined degree of similarity are clustered together by Hierarchical Agglomerative Clustering (HAC) [Zhang et al., 2005].

Blogs are becoming an important and somewhat indispensable means of information dissemination. The hidden information in these social structures has a significant impact on the rate and extent of information flow. A few studies, like [Tang et al., 2009], are exploring novel ways of measuring how the hidden social structure in these networks influences the flow of information in them. Information flow in a network is tracked by using strength of relationships. In [Tang et al., 2009] appearance of any two words

---

\(^{10}\) PieSpy Social Network Bot: http://www.jibble.org/piespy/
Chapter 2: An Overview of Social Network Extraction Systems

\(W1\) and \(W2\) repeatedly in same entries is considered as an indication of their belonging to the same topic. Experimental results obtained in [Tang et al., 2009] indicate two aspects of these networks: (a) social structures play an important role in the diffusion of “interest” topics, and (b) the authority exercised by social networks in diffusion of information contained in them is somewhat directly related to the information characteristic.

### 2.3.3.5 Online Social Networking Sites-based Social Network Extraction

Online social networking sites like Facebook\(^{11}\), twitter\(^{12}\), etc. contain huge number of user profiles containing semi-structured personal data and account for majority of Internet traffic [Alim et al., 2011]. The study in [Alim et al., 2011] extracts profile data from the deep web using the approach adopted in [Park and Barbosa, 2007]. Vector of tokens is obtained by parsing the HTML content in the data pre-processing phase. Breadth First Search is then used to traverse the specified profile webpage, in case if it has not been traversed earlier, and the extracted personal details from the traversed profile webpages are placed in a repository. Friends list and their profile addresses are extracted and inserted into the repository if they have not been stored before. An online social network graph is generated from the populated repository for analysis.

In [Salvatore et al., 2010], a web agent which mocks a real user, is used to obtain an undirected graph from explicit relationships between subscribed users of Facebook. Only those profiles which are publicly available are accessed. However, studies like [Song et al., 2010] have tried to analyse the relations within message threads in addition to explicit relationships. These latent social relationships are extracted from a social networking service by analyzing the users’ activities using a modification of frequent set mining techniques [Liu, 2006]. The hidden relationship between a set of users is

---

\(^{11}\)Facebook: https://www.facebook.com/

\(^{12}\)Twitter: https://twitter.com/
extracted by determining the occurrence as well as frequency of common terms in their writing pattern. The frequency is used as a measure of the strength of the relationship. In [Luo and Huang, 2009], social networks are constructed from contents of photo albums on Facebook using unsupervised face recognition. It first determines the owner of the album from the frequency of the most occurring image (face) in the album. If two persons appear in the same photo, an edge is added between their albums.

2.3.3.6 Multi-Source Data-based Social Network Extraction

Most of the techniques discussed above focus on a single source and the issue of social network extraction from different sources on the Web has not been discussed well in literature [Ting et al., 2009]. Combination of instant messages and e-mails for social network extraction has been proposed in [Wang et al., 2010a]. It has two major components: one for offline data collection; and the other for online data processing. Related communication data from e-mails and instant messenger is collected, the data so extracted is filtered by the data extraction engine and relevant data is stored in the database. Data collected, processed and stored by the offline data collection module is used in online processing module for social network construction and visualization.

Average time spent on each page by every user as recorded by a web server log file is the source of relationship data in [Ansari and Jalali, 2011]. The log files, after pre-processing, go to the clustering architecture. The clustering architecture consists of five sequential steps viz. site structure mining, outliers deletion, user interest discovery, user clustering and compression. The importance rate of each page in [Ansari and Jalali, 2011] is obtained based on the average time spent on each page by every user. User’s virtual communities are constructed from log files using web mining techniques.
Different dimensions of a relationship had never been explored for extraction of a social network till it was first proposed in [Kazienko et al., 2011] to analyse the social network on the basis of three dimensions viz. relations, time and groups. In [Kazienko et al., 2011] activity and interaction information was captured along with various dynamics of user’s actions. The proposed model utilizes information about different relations between various actors and the virtual groups or communities that exist based on the target relation in a particular time frame. The relations in the layer dimension describe all the relationships existing between various users of a system, both direct relationship, like e-mail or phone and indirect relationships like product recommendations, commenting of the status of a non-friend, etc. In this layered structure, each single layer uses one type of relationship information that exists between users of the system under consideration to produce a social network. Aggregation of the different layers, so obtained, results in a multi-layered social network. This social network has the same set of nodes as that of each individual layer but they are linked by multiple relations with each layer representing a relation. Group dimension contains all the social groups that can be obtained in the clustering process. The time dimension may represent a glimpse of the social network and the existing relations at any given point in time. It may also represent a set of relationships extracted based on human actions for a given period.

2.4 Name Disambiguation

Social networks try to imitate the real life relationships in a virtual world. Therefore, profile extraction/integration becomes the fundamental task, which in turn depends upon name disambiguation. It thus emerges that whatever the information source may be, recognition of entities is of great importance for extraction and analysis of social networks. The social network extraction proposed by Henry Katz [Katz et al., 1997] uses co-occurrence of names on the Web as a measure of the strength of relationship between two
entities, academics in this case. However in this case the target entities are limited and confined to a particular academic field but in real life networks, it may not be possible or even desirable to limit the target entities.

Sharing common or similar names is a common phenomenon. The nature of the problem becomes clear from the fact that around 90,000 names are shared by more than 100 million people in the United States as per the 1990 United States Census Bureau [Artiles et al., 2005], which means more than a thousand people have the same name. With ever increasing population and consequent increase in size and usage of the Internet this problem is bound to become more severe in the coming days. This problem is at full play in identification of entities (authors) in academic social network extraction, where it is further compounded by factors like multiplicity of naming conventions, usage of short names and aliases, typing mistakes, name misspellings, pseudonyms, etc. [Arif et al., 2014b].

In digital libraries, person name plays an important role as the whole process of storage, access, retrieval, and information integration revolves around person names [Tang et al., 2012]. Wrong integration, which may occur because of name ambiguity problem, may mix publications of one author with that of another with similar name or split the publications of one actual author in two or more groups. The first case is called as mixed citation (MC) problem and second case is called as split citation (SC) problem [Lee et al., 2005]. Resolving name ambiguity is thus one of the main tasks in extraction of social networks, more so in case of academic ones, and has been the focus of much of the research in SNE.

In a survey paper, Ferrira et al. [2012] discussed various representative automatic methods for author name disambiguation. Based on the type of approach used, there are two major categories as per [Ferrira et al., 2012] viz. Author Grouping and Author Assignment. Author assignment strategies are further classified into supervised techniques employing a classification method and unsupervised techniques employing a clustering technique.
Tang et al. [Tang et al., 2012] also classify these methods as being *supervised, unsupervised, and constraint-based.*

### 2.4.1 Supervised Techniques

Supervised techniques, e.g. [Han et al., 2004; Veloso et al., 2012] try to learn certain disambiguation rules for each author from human-labelled training data. In [Han et al., 2004], two supervised methods for author name disambiguation were proposed, one, based on Naïve Bayes (NB), and the second based on Support Vector Machines (SVM). Generative statistical models like NB require only positive examples in the learning process. So in the Naïve Bayes methods proposed in [Han et al., 2004] only publications of a single author were used for training the model which could later use the knowledge gained during the learning process for comparing the writing patterns of the author under consideration. Discriminative models like SVM use a classifier. In this case the model is trained by using both positive and negative examples. In this case training of the SVM is performed by using publications written by a different same name author in addition to those written by the actual author. In NB method a few citations are used for training and it is supposed that data for each of the training citation are generated by the model. In SVM the goal is to find an optimal linear decision hyperplane based on the training vectors.

The technique proposed in [Veloso et al., 2012] infers the authors of a publication by using a supervised rule-based associative classifier. Certain association rules on publication attributes are used to infer the most probable author of a given publication. During disambiguation new examples are incorporated in training data by using reliable predictions and detecting authors absent in the training data.

Peng et al. (2012) proposed a model based on Web correlations and authorship correlations using a classifier. Their method tries to resolve the ambiguity by looking for co-occurrence of two ambiguous publications on a
web source. Occurrence of two such publications is assumed as a proof of them belonging to a single author. These methods try to infer the authors of a publication by using various publication attributes like author(s), title, venue, etc. Domain specific name disambiguation is performed in [Huynh et al. 2013]. Five supervised learning methods including SVM and Random Forests are used to disambiguate Vietnamese authors only.

2.4.2 Unsupervised Techniques

The disadvantages of supervised methods prompted researchers to devise unsupervised techniques for name disambiguation which do not require any training, are flexible, have low computational costs and have been able to achieve good results [Liu et al., 2014b]. Unsupervised methods use a clustering algorithm to group citation records by clustering citation-records on the basis of similarity between selected attributes such as social network, link structure or co-authorship attributes [Wang et al., 2010b; Liu et al., 2014b]. A brief discussion about clustering methods for name disambiguation has been provided in Chapter 4.

Various name disambiguation methods [Han et al., 2005; Huang et al., 2006; Tan et al., 2006; Bhattacharya and Getoor, 2007; Culotta et al., 2007; Fang et al., 2011; Kanani et al., 2007; Shu et al., 2009; Soler, 2007; Song et al., 2007; Yang et al., 2008; Torvik and Smalheiser, 2009; Kang et al., 2009; Pereira et al., 2009; Cota et al., 2010; Tang et al., 2012; Imran et al., 2013; Liu et al., 2014a; Liu et al., 2014b] have used different available clustering techniques to overcome the inherent problems in dealing with this task to partition citation records into groups. Some of them use hierarchical agglomerative clustering, some others partitioning clustering, few spectral or graph based clustering, and a few density based clustering.

Some of these techniques obtain additional information from various sources on the Web in pursuit of improvement of the disambiguation performance. Kang et al. [2009] use the frequency of occurrence of common
co-authors in two citations on the Web as a measure for determining the similarity between two ambiguous citations in the single-link hierarchical agglomerative clustering process. Pereira et al. [2009] use Web information for disambiguating ambiguous citations in a three step process. They create a set of documents $D$ retrieved from a search engine on the basis of a query containing attributes of ambiguous citations in the first step. Next, they filter the set $D$ to obtain documents that contain publications of a given author, whereas, in the last step, they cluster citations in a hierarchical manner. Tan et al. [2006] use the publication title to query a search engine. The set of webpages returned by the search engine are treated as vectors. They introduce the concept of Inverse Host Frequency (IHF), akin to Inverse Document Frequency used in information retrieval, to assign higher weight to rare webpages which are more relevant, e.g. homepages of authors. On the basis of additional information obtained from the webpages with higher IHF, ambiguous citation-records are clustered in a hierarchical manner.

For disambiguation purposes Yang et al. [2008] use topic correlation on the Web using association rule mining. In addition to determining topics of citations from venue information they also use author homepages for disambiguation. Kanani et al. [2007] introduce the concept of web vertex by adding the additional information obtained from the Web as a new vertex in the existing graph. New edges are added from this web vertex to existing vertices showing the probability of the existing and web vertex belonging to same author.

2.4.3 Constraint-based Techniques

In constraint-based approaches [Basu et al., 2004; Zhang et al., 2007; Wang et al., 2010b] the user provided inputs, either in the beginning or in between the clustering process, act as constraints for further clustering. This approach is semi-supervised where predefined constraints are used to improve the performance of the clustering algorithm. These predefined
constraints, sometimes also called as labels, can be generated based on background knowledge, from a small quantity of labelled data, or through user-provided feedback [Wang et al., 2010b]. Five different constraints have been used by [Wang et al., 2010b]: publication in same conference, identical co-authors excluding the principal author, publications sharing same author names, publications appearing on same webpage, and two publications appearing in a set of user feedbacks. In [Basu et al., 2004], initial set of cluster centroids is estimated from the neighbourhoods on the basis of predefined constraints. At each step in the clustering process points are assigned to clusters in such a way that it minimizes both the overall distortion of the points from the cluster centroids and the number of must-link and cannot-link constraint violations. Zhang et al. [2007] propose a probabilistic model based on Hidden Markov Random Field (HMRF) in a semi-supervised fashion. Six types of constraints including similar affiliations, e-mails, co-authors, etc. are incorporated in an objective function. More the constraint violations lesser the value of objective function and fewer the chances of the publications belonging to a single author.

In addition to the above techniques some methods [Lee et al., 2005; On et al., 2005; Cota et al., 2010] use a blocking mechanism to reduce the number of candidate citation-records for comparison. The job of blocking is to partition input data on the basis of certain considerations. It improves the efficiency of disambiguation, which may employ a clustering algorithm, by reducing the number of comparisons required to find the similarity between any two citation-records. Experimental results in related fields [Fellegi and Sunter, 1969; Pasula et al., 2003] show that the blocking mechanism improves the performance considerably.

The techniques discussed above also differ in the number and type of attributes used for disambiguation. Some of them [On et al., 2005; Kang, et al., 2009] use co-author only. Others [Han et al., 2004; Han et al., 2005; Soler,
2007; Cota et al., 2010; Tang et al., 2012; Imran et al., 2013; Liu et al., 2014a; Liu et al., 2014b] use a combination of various attributes like authors, title, year of publication, publication venue, etc. available in metadata. Then there are some [Kanani et al., 2007; Yang et al., 2008; Kang et al., 2009; Pereira et al., 2009; Arif et al., 2014c] which use additional information obtained from the web to augment the metadata.

2.5 Summary

There are number of automatic social network extraction techniques proposed in literature, but to the best of our knowledge we are the first one to classify them based on the type of information source they use for relationship extraction. Using web mining techniques for social network extraction has an edge over the manual methods, because the later requires a lot of interaction with the profile owner, either directly e.g. interviews or indirectly e.g. questionnaires, for extraction of underlying relationships.

Name disambiguation is at the centre stage of these techniques. As observed during this study, the utility of the extracted networks depends on their ability to identify and extract the target relation correctly. Therefore, resolving name ambiguity is very important. In academic social networks, the target academic relationship i.e. co-authorship, has to be identified and extracted correctly. Resolving name ambiguity is much more difficult in digital citations. They suffer from mixed-citation as well as split-citation.

Solutions proposed till date for resolving name ambiguity in digital citations are either supervised or unsupervised. Supervised techniques try to learn a model for disambiguating new citations whereas unsupervised techniques try to cluster citations on the basis of similarity or difference between them. The techniques proposed so far use various publication attributes for finding similarity between them. Some of these techniques use only co-author information, whereas others use a combination of different
attributes. Majority of them use conventional attributes, whereas some use additional information like e-mail IDs, affiliations, etc.

K-means is a prominent choice among the partitioning clustering techniques. It works well when the number of candidate clusters is known in advance, or at least, we have a fair idea of the number of clusters. But in author name disambiguation it is very difficult to have a fair idea of the actual number of researchers to whom the candidate publications belong to.

We observed that hierarchical agglomerative clustering is used commonly for name disambiguation in digital citations. As a bottom-up approach the quality of clusters generated is significantly impacted by the choice of attributes for initial stage of clustering. Among all the attributes used for disambiguation in each clustering stage, an efficient solution must use the least ambiguous attribute in the beginning and the most ambiguous one in the end.

Co-authorship is commonly used as measure to disambiguate ambiguous authors but the increase in the number of authors and their associated publications has reduced its utility. As such other attributes like affiliation, e-mail-ID, etc. can be used to disambiguate them. These attributes if not available in publication metadata can be obtained from various sources on the Web or from PDFs of these publications.