Chapter 5
Words and Networks: How Reliable Are Network Data Constructed from Text Data?

Jana Diesner

Introduction

Social network data as well as the information produced or shared by network participants are prominent sources for studying reputation and authority in social media. Research studies on this topic often start with one or more network datasets and bring relevant substantive questions about socio-technical concepts such as the evolution of credibility to the data. This chapter deals with the reliability of network data itself and aims to shed some light on the following question: How reliable or accurate are network data depending on the data construction method for cases where text data are used as an input to this process? I provide a concise overview on some of the most common methods for constructing network data from text data sources, report on our findings from applying these methods to three corpora from different domains and genres, and derive implications and suggestions for theoretical and practical work.

Basically, network data can be collected or constructed in two ways: First, it might be explicitly available. For example, based on information about network participants, i.e. individuals or organizations who get represented as nodes in a graph, their connections, e.g. other social agents who they have friended or whose content they have commented on or replied to, and the content that network members provide or disseminate, such as their posts and tweets. In this case, existing application programming interfaces (APIs) and tools can be used to download and prepare these network data for analysis. For example, Facebook, Twitter, and YouTube provide such APIs, and network analysis tools such as NodeXL (Hansen et al. 2010) and ConText (Diesner et al. 2013) provide respective data import options.

Alternatively, network data can be constructed or inferred from textual data and metadata that are generated, authored, or disseminated by network participants. These data typically occur in the form of semi-structured or unstructured natural language text data (Corman et al. 2002; Danowski 1993; Diesner and Carley 2005).

J. Diesner (✉)
GSLIS/UIUC, 501 E Daniel Str, 61820 Champaign, IL, USA
e-mail: jdiesner@illinois.edu

E. Bertino, S. A. Matei (eds.), *Roles, Trust, and Reputation in Social Media Knowledge Markets,* Computational Social Sciences, DOI 10.1007/978-3-319-05467-4_5,

In computing, this process is also known as relation extraction (Bunescu and Mooney 2005; Culotta et al. 2006; Roth and Yih 2002).

Besides distilling network data from text data, text data can also be used to enhance explicitly given social network data with the information authored or disseminated by network members. This can be done, for instance, by linking nodes representing agents to nodes representing highly salient information associated with these agents. The resulting networks are typically referred to as socio-semantic networks (Diesner 2012; Gloor and Zhao 2006; Roth and Cointet 2010). One of the main advantages of considering text data for network analysis is that this approach allows for studying the interplay and coevolution of information and social networks. This includes the transformative role that language can play in networks and vice versa (Milroy 1987).

Overall, constructing or enhancing network data based on text data involves a plethora of decisions that have to be made. For example, how to identify nodes and linking them into edges. These decisions can majorly impact the understanding that end users gain about a network and any conclusions they draw from that. The problem here is that the impact of these choices on the resulting relational data is insufficiently understood. This chapter focuses on the different views of a network that one can get when using different relation extraction methods. Who cares about this knowledge? I argue that an empirically grounded understanding of the impact of choices made for text analysis on the derived networks structures contributes to an improved comparability and generalizability of respective methods and tools. Furthermore, such knowledge helps researchers and practitioners to draw valid and reasonable conclusions from analysis results. This is particularly important in cases when validating network data against ground truth data is hard to infeasible, e.g. in the case of covert or historic networks.

From Words to Networks: Methods for Constructing Network Data from Text Data

Network Construction Based on Text Data

In the (computational) social sciences and (digital) humanities, textual data are often converted or coded into networks by developing and applying a codebook (Abello et al. 2012; Gerner et al. 1994; Roberts 1997). Codebooks contain rules for translating relevant pieces of text data into code. These codes represent relevant categories for studying a certain topic, domain, or corpus. Applicable categories can be identified in a top-down fashion from theory and/or in a bottom-up or empirical fashion from the underlying data (Bernard and Ryan 1998; Glaser and Strauss 1967). Node classes can also serve as codes, e.g. “agents”, “organizations” and “locations” (Diesner and Carley 2008). Multi-columned tables that associate text terms with codes are also referred to as thesauri or dictionaries. Traditionally, codebooks and thesauri were created in a manual or semi-automated fashion (Bernard and Ryan 1998), which

allows for incorporating human expertise, manual verification of the term to code assignments, and the creation of a controlled vocabulary at the cost of scalability and generalizability (Diesner 2012). Alternatively, techniques from natural language processing and/or machine learning can be applied to create codebooks and thesauri (Cohen and Sarawagi 2004; Diesner and Carley 2008; Roth and Yih 2002), which enable the efficient coding of vast amounts of text data sources (Abello et al. 2012).

The identified instances of relevant entity classes can further be used as nodes for constructing networks. Common approaches for linking nodes into edges rely on (a mixture of) co-occurrence or proximity as well as semantic, syntactic, and statistical features of the text data. While proximity-based approaches have been criticized for their arbitrariness (Corman et al. 2002) and potentially high ratio of false positives (Diesner 2012), it is the most common technique for linking codes or nodes into edges. Technically speaking, proximity-based node linkages result in association networks; a very common type of relational structures extracted from text data. The considered node classes determine the type of network that gets constructed: for example, when identifying social agents (people and organizations) from text data, the resulting graphs represent social networks. When retrieving instances of knowledge and information and the connections between them, the resulting networks can represent semantic networks (Diesner and Carley 2011; Woods 1975). We are taking a more humble approach herein by referring to networks where nodes represent instances of knowledge and information referenced in the text data as knowledge networks.

Network Construction Based on Metadata

While codebook applications operate on the content level, metadata associated with text corpora can serve as another or supplemental source of information for constructing network data. For example, when using LexisNexis—a provider of large collections of data from various sources and genres—to search for documents, the retrieved articles can be downloaded along with metadata. These metadata concisely index the content of the underlying text bodies along various categories. For the case of news wire data, for example, these categories entail "person" and "organization" (social agents), "geographic" (locations), and "subject" (themes). Furthermore, in LexisNexis, each metadata entry is associated with a relevance score that indicates the strength of the association of an article with an index term. Resembling the idea of proximity-based link formation as discussed above, indexed keywords can be linked into edges if they co-occur for the same article. The link weight can be increased accordingly when the same pair of index terms is observed for multiple articles. Another prominent source for metadata are keywords for research proposals and publications that authors select when submitting a paper. Such keywords can be based on a predefined catalogue of eligible terms (controlled vocabulary) and/or identified by the authors given the content of their documents.

Building (multi-modal) network data from metadata is a highly efficient process: Once the metadata are organized, e.g. in a database, the network construction process becomes basically a search and retrieval routine. The ConText software for example supports the construction of metadata databases from previously downloaded LexisNexis files, and the construction of one- and multi-modal network data from these databases (Diesner et al. 2013). The limitation with this approach is that the assignment of metadata entries and relevance scores to articles is not always transparent. For LexisNexis, for example, there is no publicly available documentation on the algorithms or methods used for this process.

Ground Truth Network Data

One way to assess the accuracy of relation extraction techniques and network construction based on text data and metadata is to compare the obtained results against ground truth data, which are also referred to as gold standard data. Ground truth data are typically generated by humans who are specifically trained for this task. Humans can construct ground truth network data in two ways: first, by performing relation extraction based on some text corpus by hand, typically in a computer-supported fashion, and second, by denoting network data to the best of their knowledge, generally also in some computer-assisted way. Both processes are assumed to result in reliable or validated data at the expense of costs and scalability. In other words, given the time-consuming nature of this process, it is often not possible to generate ground truth data for a large-scale dataset or networks. This fact hinders the validation of relation extraction techniques, including the evaluation of the performance of prediction models beyond accuracy rates (Diesner 2012).

Overall, the whole process of going from texts to networks and validating the resulting data is only needed or applicable if one cannot ask network members directly about their relationships or their views of a network (Krackhardt 1987). This applies, for instance, to the case of hidden or historic networks (Diesner and Carley 2005; Sparrow 1991).

Problem Statements

Given these different approaches to network construction and validation, the following research questions with high impact for practical applications are eminent yet heavily under-researched: First, given a corpus, how closely do the results from various content-based and metadata-based network construction techniques resemble ground truth data? And second, how do the outcomes of these methods compare to each other? In other words, what different views of a network do we gain when choosing one method over another? We have conducted several of these comparisons in a series of empirical experiments and report on our findings in the results section (Diesner 2012).

Data

We used three datasets for our analyses: First, our curated version of the Enron email dataset (herein referred to as Enron). This particular version contains 58,266 emails from employees of the former Enron corporation (Diesner et al. 2005). Second, a corpus of news articles about the Sudan (herein referred to as Sudan). This corpus is a curated collection of 79,388 news wire articles released between 2003 and 2008 about the Sudan (Diesner 2012). We collected these data from LexisNexis. Third, a corpus of 55,972 proposals accepted for funding through the European Framework Programmes between 1988 and 2010 (herein referred to as Funding) (Diesner 2012).

While these datasets differ with respect to genre (social media, news articles, scientific writing), domain (business, politics, science), target audience (from internal or private to public), and time span, they are comparable in that they entail text bodies plus metadata: For Enron, we used the email bodies and social agents denoted in the email headers. For Sudan, we worked with the content of the articles and the index terms assigned by LexisNexis. For Funding, we used the project title plus description and predefined index terms selected by the people who submitted the proposals. For details on these data see also Diesner (2012).

Methods

For extracting network data from text data, we built codebooks and thesauri, applied them to the text data, and linked any matches based on their proximity (for details see Diesner 2012). For each dataset, two different thesauri were constructed, which enables the comparison of the impact of different approaches to this step:

First, we used text mining techniques to identify salient terms, e.g. based on (weighted) term frequency metrics, and leveraged existing external and internal dictionaries. We manually verified, consolidated, and disambiguated every entry. This process took between two days (Enron) and six weeks (Sudan) where the time costs mainly depend on the quality and compatibility of leveraged existing material. I refer to this process as relation extraction based on classic codebook construction (CCC). For all three datasets, we aggregated networks per year (Sudan), funding period (funding) and stages of the organizational crisis (Enron) into cumulative graphs per time chunk. The same procedure was also used for the next two methods.

Second, we ran prediction models for entity extraction on each corpus. I refer to this process as entity extraction-based codebook construction (EECC). We had built these models by using conditional random fields, a supervised machine learning technique particularly suited for learning from sparse, sequential data where it is highly beneficial to exploit long-range dependencies (Diesner 2012). Our models go beyond the classic set of named entities (people, organizations, locations) by also detecting other entity classes that are relevant for modeling socio-technical systems, such as resources, tasks, events, knowledge, and attributes, as well as instances of entities that are referred to by a name (e.g. Barack Obama) or not (e.g. politician).

While the models achieved accuracy rates (F scores) of 87.5–88.8 % during the k-fold cross-validation of the machine learning process, applying them to our datasets and again manually verifying their fitness showed that thesauri built this way also need some post-processing in the form of reference resolution and cleaning. Still, the EECC approach outperforms the alternative CCC process in terms of time costs, with this process taking seconds to a few minutes for generating a thesaurus per corpus and up to 2 days for post-processing it. Moreover, the prediction models generalize with known accuracy while a thesaurus built in the classic way for one dataset cannot be assumed to generalize well to corpora from other domains, genres, or points in time due to the deterministic nature of thesauri.

For constructing metadata networks, for Sudan, we linked any two entities occurring in the metadata that represent people, organizations, locations, or knowledge per article into bidirectional, weighted graphs. The weights were identified by computing the average of the lowest-relevance scores for any two linked entities. For Funding, we coded all index terms as knowledge and linked any such pairs per proposal into edges. For Enron, we connected senders and receivers (to, cc, bcc) into directed social networks that were weighted by the cumulative frequency per entity pair. Note that this approach defines a classic, explicitly given social network; the way it is often constructed from social media data. The resulting network can then be compared against social networks extracted from the text data. Each of these operations was a matter of minutes once we had curated the data and organized them in relational databases.

As for ground truth networks generated by human experts, we were only able to construct such data for Sudan. This was possible through a collaboration with Dr. Richard Lobban, a leading expert on the Sudan, and his team. More specifically, we went through a qualitative, computer-supported, iterative process of building expert-verified networks of tribal affiliations in the Sudan for each calendar year of 2003–2008. We started by applying a list of all tribes in the Sudan, which was provided by Dr. Lobban's team, to our Sudan corpus, creating a first visualization of the tribal network and sending that to Dr. Lobban for verification, i.e. annotating false positives and false negatives in terms of nodes and edges. Once we received their modified maps, we adjusted our coding scheme and regenerated the network data. We repeated this process until Dr. Lobban's teams assessed the networks as representative of the ground truth based on their expertise. The time costs for this process are comparable to building codebooks without leveraging machine learning methods. Since this process cannot be expected to scale up, it can only be used for small to moderately sized networks.

Once we had constructed these networks, we compared them within and across datasets and methods. More specifically, we identified the structural overlap of nodes based on their node names or labels and the edges between them.

Results and Conclusions

How much do network data constructed from text data or metadata resemble ground truth data? It depends, but overall very little, as our results suggest: Out of the social

network data built in collaboration with subject matter experts, 53 % of the nodes and 20 % of the links also appeared in networks distilled from text bodies when using the classic thesaurus construction (CCC) approach. These values drop to 11 % for nodes and 5 % for edges for relation extraction based on automatically built thesauri (EECC), and to flat zeros for metadata-based networks. What accounts for these differences? The main reason for the overlap between networks based on ground truth data and the CCC method is that we reused the same list of tribes as input for the human experts and thesaurus construction while the EECC method finds any applicable matches purely based on the underlying machine learning techniques. We observed the same effect for other methods: CCC-based networks resemble metadata networks more closely than the EECC-based networks, mainly because we enhanced the classic thesauri with data that we also used for defining nodes for metadata networks, e.g. lists of index terms. At the same time, EECC-based networks and metadata networks are constructed from different data, namely text bodies and metadata; with the differences in terminology and scope leading to different views of the networks. In summary, reconstructing social networks by applying text mining techniques to corpora, including metadata, will lead to largely incomplete and biased incomplete results. This limitation could be alleviated by switching from proximity-based node linkage to alternative methods, such as approaches based on syntax, semantics, and machine learning techniques (Roth and Yih 2002; Zelenko et al. 2003). In our studies, the structural agreement between any pair of networks was consistently higher on the node level than on the edge level. This effect might also change by using different node linkage strategies.

Another factor that we observed to strongly impact the agreement in networks structure is the network size: Larger networks have a higher chance to resemble parts of networks constructed with alternative methods that lead to smaller networks, both in number of nodes and edges. This fact is of methodological and practical relevance since various network metrics have been shown to correlate with network size (Anderson et al. 1999; Friedkin 1981).

Comparing networks not on a structural but a substantive level leads to different findings depending on the domain and network construction method: For corpora of news articles, social networks created from metadata feature major international key entities and their connections, while social networks distilled from text bodies provide to a more fine-grained and localized understanding of important actors and their links. In contrast to that, when looking at knowledge networks across the genres and domains considered, text-based networks give a high-level overview on salient terms and their connections for a given domain, while metadata networks drill down to more specific pieces of knowledge and information per domain. The reason for this effect is that the keywords and index terms from which the metadata networks are constructed are already highly condensed and often carefully selected mini-summaries of the underlying text bodies while these concepts are being elaborated on in a more detailed fashion in the actual text bodies. This explanation also partially accounts for another observation: Looking at ambiguity issues in the generated networks, we found that metadata networks are less limited by co-reference resolution issues than methods that operate on the content level. Co-reference resolution here means disambiguating terms with the same surface form but different meaning and consolidating terms with different surface forms but the same meaning.

Synthesizing our findings, we recommend fusing text-based and metadata-based networks in an informed fashion: using machine-learning-based entity extraction to build a thesaurus, refining and enhancing it based on subject matter expertise if available, and linking up nodes based on methods that are more advanced than co-occurrence is best suited for generating social networks. These networks can be combined with knowledge networks derived from metadata. Based on our findings, the resulting networks will allow for a broad and deep look into social and knowledge networks.

Acknowledgements This work was supported by the National Science Foundation (NSF) IGERT 9972762, the Army Research Institute (ARI) W91WAW07C0063, the Army Research Laboratory (ARL/CTA) DAAD19-01-2-0009, the Air Force Office of Scientific Research (AFOSR) MURI FA9550-05-1-0388 and the Office of Naval Research (ONR) MURI N00014-08-11186. Additional support was provided by CASOS, Carnegie Mellon University. The views and conclusions contained in this chapter are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the NSF, ARI, ARL, AFOSR, ONR, or the United States Government.

References

Abello, J., Broadwell, P., & Tangherlini, T. R. (2012). Computational folkloristics. *Communications of the ACM, 55*(7), 60–70.

Anderson, B. S., Butts, C., & Carley, K. M. (1999). The interaction of size and density with graph-level indices. *Social Networks, 21*(3), 239–268.

Bernard, H., & Ryan, G. (1998). Text analysis: Qualitative and quantitative methods. In H. Bernard (Ed.), *Handbook of methods in cultural anthropology* (pp. 595–646). Walnut Creek: Altamire Press.

Bunescu, R., & Mooney, R. (2005). Subsequence kernels for relation extraction. *Advances in Neural Information Processing Systems* (NIPS), Vancouver, BC.

Cohen, W. W., & Sarawagi, S. (2004). *Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods*. Paper presented at the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA.

Corman, S. R., Kuhn, T., McPhee, R. D., & Dooley, K. J. (2002). Studying complex discursive systems: Centering resonance analysis of communication. *Human Communication Research, 28*(2), 157–206.

Culotta, A., McCallum, A., & Betz, J. (2006). *Integrating probabilistic extraction models and data mining to discover relations and patterns in text*. Paper presented at the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. New York, NY.

Danowski, J. A. (1993). Network analysis of message content. *Progress in Communication Sciences, 12,* 198–221.

Diesner, J. (2012). *Uncovering and managing the impact of methodological choices for the computational construction of socio-technical networks from texts*. Carnegie Mellon University. (CMU-ISR-12–101, PhD Thesis).

Diesner, J., & Carley, K. M. (2005). Revealing social structure from texts: Meta-Matrix text analysis as a novel method for network text analysis. In V. K. Narayanan & D. J. Armstrong (Eds.), *Causal mapping for information systems and technology research: Approaches, advances, and illustrations* (pp. 81–108). Harrisburg: Idea Group Publishing.

Diesner, J., & Carley, K. M. (2008). Conditional random fields for entity extraction and ontological text coding. *Journal of Computational and Mathematical Organization Theory, 14,* 248–262.

Diesner, J., & Carley, K. M. (2011). Semantic networks. In G. Barnett & J. G. Golson (Eds.), *Encyclopedia of social networking* (pp. 766–769). Sage.

Diesner, J., Frantz, T. L., & Carley, K. M. (2005). Communication networks from the Enron Email Corpus. It's always about the people. Enron is no different. *Computational Mathematical Organization Theory, 11*(3), 201–228.

Diesner, J., Aleyasen, A., Kim, J., Mishra, S., & Soltani, S. (2013). *Using socio-semantic network analysis for assessing the impact of documentaries*. Paper presented at the WIN (Workshop on Information in Networks). New York, NY.

Friedkin, N. E. (1981). The development of structure in random networks: An analysis of the effects of increasing network density on five measures of structure. *Social Networks, 3*(1), 41–52.

Gerner, D., Schrodt, P., Francisco, R., & Weddle, J. (1994). Machine coding of event data using regional and international sources. *International Studies Quarterly, 38*(1), 91–119.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. New York: Aldine.

Gloor, P., & Zhao, Y. (July 2006). *Analyzing actors and their discussion topics by semantic social network analysis*. Paper presented at the 10th IEEE International Conference on Information Visualisation. London, UK.

Hansen, D., Shneiderman, B., & Smith, M. A. (2010). *Analyzing social media networks with NodeXL*. Morgan Kaufmann.

Krackhardt, D. (1987). Cognitive social structures. *Social Networks, 9*(2), 109–134.

Milroy, L. (1987). *Language and social networks* (2nd ed.). Oxford: Blackwell.

Roberts, C. W. (1997). *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*. Mahwah: Lawrence Erlbaum Associates, Inc.

Roth, C., & Cointet, J. (2010). Social and semantic coevolution in knowledge networks. *Social Networks, 32*(1), 16–29.

Roth, D., & Yih, W. (2002). *Probabilistic reasoning for entity and relation recognition*. Paper presented at the International Conference on Computational Linguistics (COLING), Taipei, Taiwan.

Sparrow, M. (1991). The application of network analysis to criminal intelligence: An assessment of the prospects. *Social Networks, 13*(3), 251–274.

Woods, W. (1975). What's in a link: Foundations for semantic networks. In D. Bobrow & A. Collins (Eds.), *Representation and understanding: Studies in cognitive science* (pp. 35–82). New York: Academic Press.

Zelenko, D., Aone, C., & Richardella, A. (2003). Kernel methods for relation extraction. *The Journal of Machine Learning Research, 3,* 1083–1106.