Social network visualization from TEI data

Marcus Bingenheimer, Jen-Jou Hung and Simon Wiles
Library and Information Center, Dharma Drum Buddhist College, Taiwan, R.O.C.

Abstract

The focus of this article is a system for visualizing social network data derived from a TEI-encoded corpus of texts. It describes the collection of biographies of historical Chinese Buddhist monks, which constitutes this corpus and the TEI markup, in particular the innovative concept of a 'nexus-point' that was originally applied to them with the goal of producing GIS-like visualizations [see Bingenheimer, M., Hung, J.-J., and Wiles, S. (2009). Markup meets GIS - Visualizing the 'Biographies of Eminent Buddhist Monks'. In Banissi, E. *et al.* (eds), *Proceedings of Information Visualization IV 2009*. IEEE Computer Society: 550–4.]. Over the course of this work, it became clear that a data set of nexus-points could be derived from this markup which would support a representation of the social network which can be inferred from the corpus. The nature of this social network is explored and some interesting preliminary applications are suggested. The software architecture which supports the visualization, based on the PREFUSE toolkit, is introduced. Finally, the scope for the future development of the corpus and the system are discussed, and some avenues for potentially fruitful analysis are suggested. Throughout the article, it is argued that the methods and techniques employed here are applicable well beyond the present context. In describing this project of social network visualization, it is demonstrated that a well-marked-up TEI corpus can, with very little additional technical overhead and using the same markup, serve as the basis for multiple representations of the same data.

Correspondence:
Marcus Bingenheimer, Dharma Drum Buddhist College, No. 2-6 Xishihu, Jinshan 20842, Taipei County, Taiwan, ROC.
E-mail:
m.bingenheimer@gmail.com

1 Introduction

One of the main advantages of well-marked-up texts, and the data sets which can be derived from them, is that their application is never limited to any single research question or mode of visualization. Although most projects deploy markup with particular analytic or display purposes in mind, it is often the case that the understanding encoded into the text in this way can be used in the service of multiple ends, and often in unexpected and serendipitous ways. This article introduces a way of visualizing social networks extracted from a TEI-encoded corpus whose markup conventions were originally intended to support spacio-temporal analysis (Bingenheimer *et al.*, 2009). After the markup was added, it became apparent that the data now available also describes and constitutes a social network and could be visualized as such, thereby giving rise to completely new views and perspectives on the information contained in the texts. The results are of substantial interest and have potentially great significance for research into Chinese Buddhist history. An online interface giving

Literary and Linguistic Computing, Vol. 26, No. 3, 2011.

doi:10.1093/llc/fqr020 Advance Access published on 22 May 2011

access to visualizations of the social networks contained in the marked-up corpus, built using open-source software, has been made available to scholars and the general public. We aim to show how social network visualizations can be achieved from high-end digital editions with comparatively little overhead.

```
<linkGrp>
  <link targets="#0710b22 #0710b23"/>
  <ptr type="person" target="#A001470"/>
  <ptr type="person" target="#A001606"/>
  <ptr type="place" target="#CN0410311T08AA"/>
  <ptr type="time" target="#d54866395486668"/>
</linkGrp>
```

Fig. 1 Nexus-point attached to a biography: several persons at a certain place at a certain time

2 Data set

Between 2007 and 2010, Dharma Drum Buddhist College conducted a markup project with the aim of creating advanced digital editions of four important collections of biographies of Buddhist monks (the *Liang-* 梁-, *Tang-* 唐-, *Song-* 宋-, and *Ming-* 明- *Gaoseng Zhuan*高僧傳). More than 1,300 biographies were marked up and each place name, person name, and date was connected to an authority database. The corpus contains information on more than 5,900 persons and more than 3,500 places.

2.1 Nexus-points

In addition to identifying and disambiguating these entities, geo-referencing the places, and providing basic information on the people, another layer of markup was added to record semantically simple events that joins these data. We call this markup construct a 'nexus-point' because it represents information in the text which connects people, places, and time. Nexus-points in the data set described here can be expressed in natural language in the form:

> One or more persons were at a certain place at a certain time.

In the markup, this was realized using link groups connecting persons, places and dates that are mentioned in a biography. A reference to the exact location of the nexus-point in the text (a TARGET attribute on <link> elements) allows easy look-up in case of longer biographies (see Figure 1).

Nexus-points can of course be expressed in formats other than TEI (such as RDF, OWL, etc.) but they were originally conceived by us in the context of a markup project that was to serve as the basis for a spacio-temporal exploration of Chinese Buddhist history. The main characteristic of nexus-points is that they are syntactic, not semantic—i.e. nexus-points make no statement as regards exactly what happened with these person(s) at a certain time at a certain place, but only that the text contains information that something happened. This is a great advantage. Firstly, because it does away with the need to devise an ontology of events. Secondly, it does not preclude future extension of the data. Should researchers later want to add categories for these events, they can do so by building on the indispensable foundation of clearly identified objects (persons, places, dates, etc.).

For this data set, the date given for events is only rarely a particular day or month, but often points to a time period. In some cases, especially when short biographies do not contain an explicit place-date-person nexus, we use the life dates of a person as the date reference. Such a nexus-point is equivalent to saying that the person was at a certain place at some point in his or her life. This is still useful for a geospacial visualization that shows activity, or the absence of activity, in certain locations in China at different periods. For our social-network visualization, the lack of precision for certain events does not impact their usefulness. Our aim is to show 'who knew whom' in a certain period of Chinese Buddhism. Where the texts give precise dates, they are reflected in the markup; where the texts are imprecise, we record the range of possible values, the reign period of an emperor, or the life dates of a person.

For the visualization of social networks, we can use only those nexus-points that connect more than one person, thus yielding at minimum dyads of two

actors which are not further connected to others. As of August 2010, the corpus contains 5,565 nexus-points that are used for the visualization of geospacial and temporal patterns, and a subset of 2,126 of these connect more than one person, making statements of the form:

> Two or more persons were at a certain place at a certain time.

The data set of these 2,126 events is the basis for our first attempt at visualizing social networks.

3 Social Network

Historical social network analysis has been used sporadically for decades.[1] Our own approach differs from previous attempts to use social network tools with historical data in that it is essentially text based and not centered on a particular research question. Furthermore, while traditionally social network visualization is more often a result of an analytical procedure, the aim here is to show that social networks can be 'visualized' from marked-up texts almost as a by-product: the information extracted is valid but not biased toward any particular use, because only very basic parameters—personal identity, place, and time—are encoded. Our focus is on producing useful visualizations of high-end digital editions of classical Buddhist texts. As such they are the beginning of further inquiry rather than the result of historical analysis.

The biographies of eminent monks are usually short, mostly comprising only a few paragraphs in English. They are generally told chronologically following the career of an eminent monk from his childhood, to where he received his education. After that, important events or even just a single story are mentioned ('was summoned by the emperor,' 'translated sutra *X* together with the Indian master *Y*,' 'went to place *Z* and taught Vinaya there', etc.), again in chronological order, and the biography closes with the time and circumstances of the subject's death (where known). Buddhist biographies in East Asia follow the genre established by Chinese historiography in the second century BCE which, in marked contrast to the Indian model, tries to give time and place information wherever possible. It is in the nature of these texts that wherever two or more persons are mentioned as being in the same place at the same time, these people almost certainly knew each other. Actors which are not somehow interacting with either each other or the main subject are simply not part of the narrative.

As mentioned above, nexus-points do not contain information as to the nature of the event: they are devoid of semantic content. To include information about what happened would require an ontology of events, the development and deployment of which would have slowed down the markup process considerably. Moreover, the aim of this project was to provide a general purpose corpus for an advanced representation of biographical literature, and an ontology would almost necessarily be biased toward a particular research agenda: 'The choice of relational content, also called *type of tie*, is largely determined by a project's theoretical concerns and research objectives' (Knoke and Yang, 2008). However, the corpus was designed to be extensible, and should a particular ontology of events be needed one day, it can now be easily added by extending the link groups that connect two or more actors.

At the present time, nexus-points express only the most basic social connection between actors: that they communicated with each other in some way.[2]

As of August 2010, the largest simple visualization that we can draw from our data set contains 2,467 actors connected by 4,711 ties. Average connections per actor are 2.26 and actual number of connections range from one to eight. The density of the network is 0.0008.

Figure 2 shows the largest possible social network derived from our data (using PREFUSE; see below). The periphery shows a belt of small networks, with a number of unconnected dyads at the outer fringe. These are actors that are only mentioned in connection with one other actor, who in turn has no other ties. In the center, there is a large cluster where all actors are connected with each other via walks of varying length. Subgroups and cliques (groups where all actors know each other) are already

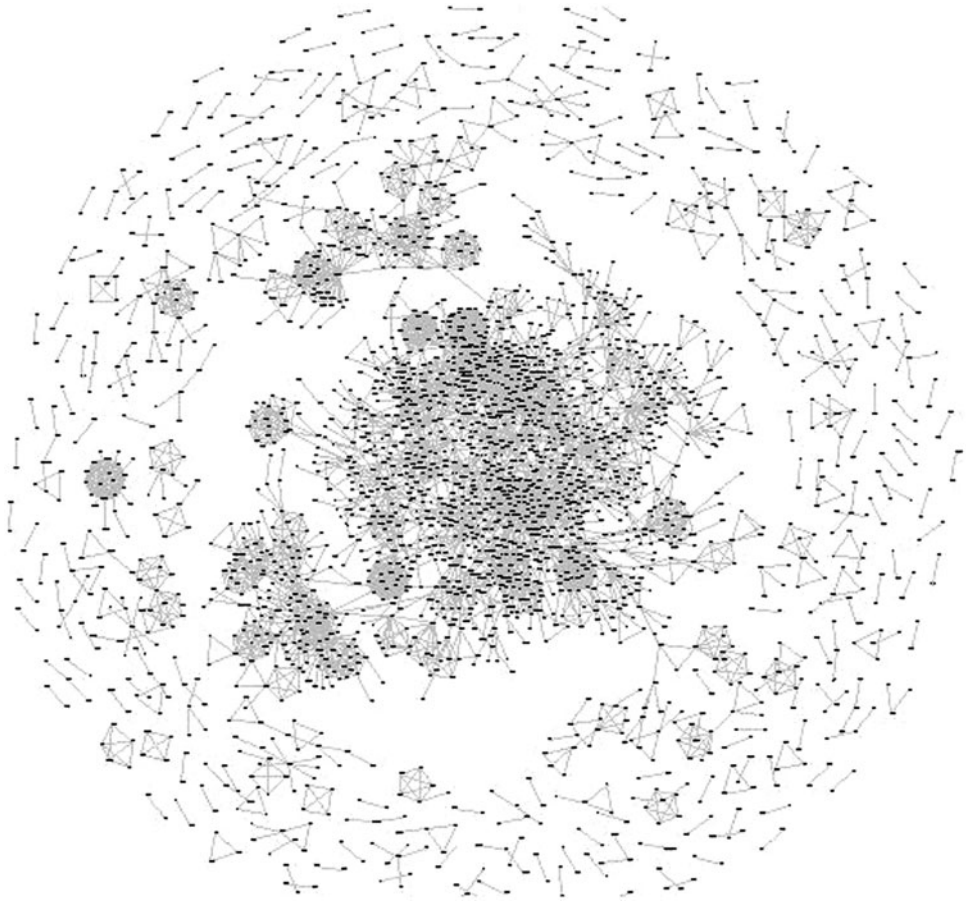

Fig. 2 Visualization of the largest data set: all actors, all ties

recognizable and become increasingly evident on closer inspection.

The simplicity of the nexus-point events means that all graphs drawn from this data set are undirected. Ties between actors could be weighed by counting the number of nexus-points that establish this particular connection. Stronger ties point to more information about interaction between individuals. Another way to valuate ties would be to analyze the network according to the categories in which the biographies are grouped in GSZ literature (translators, exegetes, Vinaya masters, thaumaturges, etc.). Though these would be only applicable for monks, it would still be interesting to see if different categories evince different network structures.

4 Use of Social Network Visualization

Our short-term aim is to produce an online interface to visualize the communication network of Chinese Buddhist history, as it can be learned from the sources, as a research tool for scholars. Already in the trial version, the tool answers the question 'Who knew whom (according to our data set)?,' a type of information that was not available previously through dictionaries or indices. Users can easily identify cliques (groups of actors who know each other), prominent actors (indicated by larger labels), and multiple ties between individuals (indicated by thicker lines).

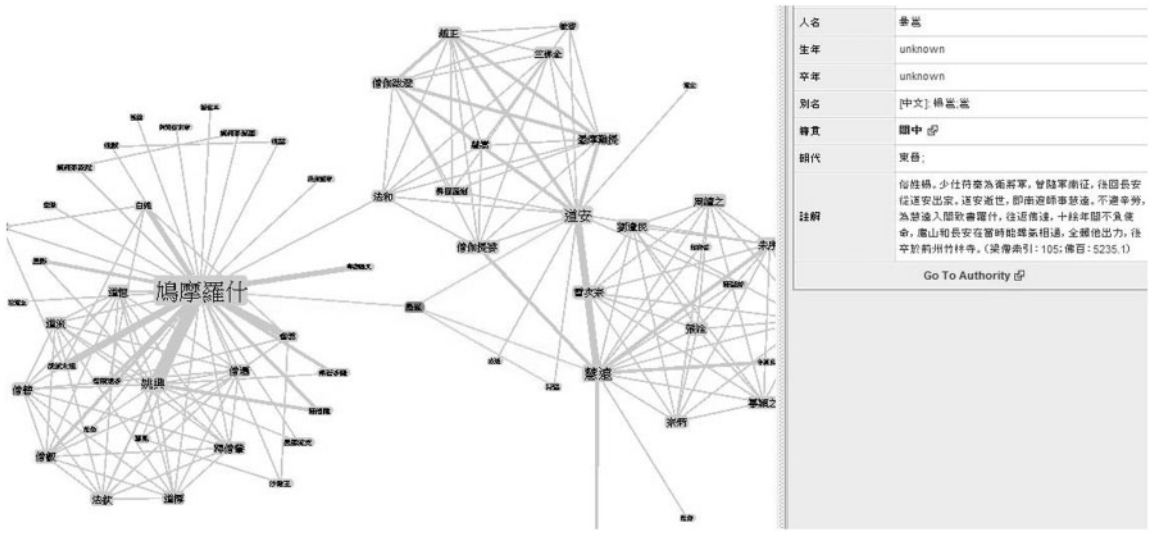

Fig. 3 One actor serving as a bridge between different subgroups

In the example shown in Fig. 3, the monk Tanyong 曇邕 (fl. late fourth century CE) connects the sub-systems of three key figures in the history of Chinese Buddhism. In the current data set, the famous translators Daoan 道安, Huiyuan 慧遠 and Kumārajīva and their respective cliques are connected by only one single person: Tanyong. Tanyong is decidedly not a famous person and hardly appears in any modern study of Buddhist history, though his role in the communication flow of the Buddhist elite of those days certainly deserves mention.

We are not yet at the stage where we can confidently use the data set for sophisticated social network analysis, but simple analytic procedures which identify actors that connect cliques or groups or actors with a high degree of centrality are well within our reach.

5 Visualization Interface Architecture

Our interface is currently built on PREFUSE (http://prefuse.org/), a Java-based visualization toolkit which yielded good results within a short development time, and runs as an applet in a user's browser.

The strengths of PREFUSE are its high quality built-in data visualization models and its powerful functions for user interaction. These features are packaged in simple and easy-to-use components that allow the construction of a professional visualization system with minimal effort. The logical structure of the PREFUSE library is organized according to the Information Visualization Reference Model (Chi *et al.*, 1999), a well-known information visualization framework, which ensures that PREFUSE is easy to integrate and work with. PREFUSE is an open-source software, meaning that the option is always available to modify its internal functionality if necessary. Comprehensive documentation for PREFUSE is available, which is not always the case with open-source tools.

PREFUSE has many different visualization models, which it refers to as 'layouts', including 'AxisLayout', 'CircleLayout', 'GridLayout', and TreeLayout. Each layout has unique features and therefore is suitable for different visualization scenarios. Our visualization system is based on 'ForceDirectedLayout', a specialized graph-like layout in which the visualization basically

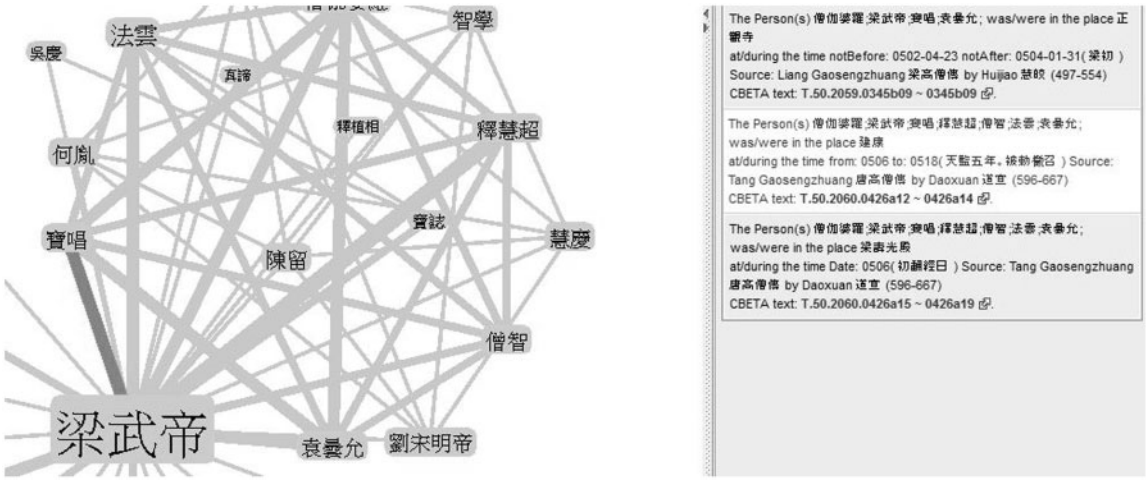

Fig. 4 More nexus-point information results in thicker ties, links connect to the corpus

consists of a set of nodes and edges binding related nodes. In contrast to a standard graph layout, 'ForceDirectedLayout' adds repellent force between nodes, and makes edges act as springs. This modification tightens the bonds between nodes in the same cluster and increases the distance between clusters, such that the boundaries between node clusters in the social network are more prominent. In addition, we also overwrite the default 'ColorAction' class to highlight nodes that are directly connected to the currently active node.

The interface provides query filters for persons, time and biographical collection. By applying queries such as 'Show me information on social networks for the time between 450 and 500 CE', the current interface allows for visualization of partial networks.

The interface is connected to our authority databases and clicking on a person label yields basic biographical information on the actor (see Fig. 3).

Clicking on a tie, one gets information on the one or more nexus-points that contain the information on which the tie is based (see Fig. 4). From there, links back to the biographies allow users to verify the nexus-point event in the original texts.

6 Future Development

The nexus-point concept has proved a useful vehicle for representing textual information, and we have managed to produce tools for GIS and social network visualizations. Its application is obviously not limited to Buddhist Studies but can be used wherever one deals with complex, linguistically difficult sources. Through the identification of persons, places, and dates and the joining of these three data to create nexus-points, a significant amount of information is made available for computation. Both the tools and the data set have begun to be used by scholars in Buddhist studies as well as other fields.[3]

In the near future, we will encode further information by defining new types of nexus-points, relaxing the stringent requirement for the presence of location and date elements in a nexus-point. Type II nexus-points can omit a location reference. Type II events can not be used for GIS-like visualizations, but contribute to the data set available for visualization of social networks. We are also in the process of adding more biographical collections to our corpus of texts.[4]

The online visualization tool, useful as it is to understand the social connections of a monk

according to the current data set, is only a first step toward further analysis on the level of the whole network. As Wetherell (1999) remarks 'Because historians are plagued by an incomplete historical record and imperfect understandings of past social relations, HSNA [Historical Social Network Analysis] remains an inherently problematic enterprise. Yet despite conceptual, methodological and evidentiary obstacles, SNA possesses real potential for historical analysis.' In our case, two directions of analysis are especially promising in terms of yielding new insights into Buddhist history. First, an analysis of various types of centrality is likely to identify the central players in the network described by our data. All three main types of centrality (Wasserman and Faust, 1994)—degree, closeness, and betweenness—have the potential to contribute to our understanding of Buddhist history. Second, algorithmic analysis of cliques appears promising, because it allows us to identify groups of actors who knew each other. Working on any one single actor X in history, it is useful to know if the people he or she interacted with also knew each other.

Funding

This work was supported by the Haoran Foundation 浩然基金會.

References

Bingenheimer, M., Hung, J.-J., and Wiles, S. (2009). Markup meets GIS - Visualizing *the 'Biographies of Eminent Bu*ddhist Monks'. In Banissi, E. *et al.* (eds), *Proceedings of Information Visualization IV 2009*. Piscataway/NJ: IEEE Computer Society, pp. 550–4.

Carrington, P. J., Scott, J., and Wasserman, S. (2005). *Models and Methods in Social Network Analysis*. New York: Cambridge University Press.

Chi, E. H., Chi, H-H., and Riedl, J. T. (1999). *A Framework for Information Visualization Spreadsheets.* Ph.D. thesis, University of Minnesota.

Gould, R. V. (2003). Uses of network tools in comparative historical research. In Mahoney, J. and Rueschemeyer, D. (eds), *Comparative Historical Analysis in the Social Sciences*. Cambridge: Cambridge University Press, pp. 241–69.

Habermas, J. (1988). *Theorie des kommunikativen Handelns*, 2 Vols. Frankfurt/M: Suhrkamp.

Kenderdine, S. (2011). Cultural data sculpting: omnispatial visualization for large scale heterogeneous datasets. In Bearman, D. and Trant, J. (eds), *Museums and the Web, Selected Papers from Museums and the Web 2011*. Philadelphia: Archives & Museum Informatics.

Knoke, D. and Yang, S. (2008). *Social Network Analysis*. Los Angeles, London: Sage Publications.

Luhmann, N. (1997). *Die Gesellschaft der Gesellschaft*, 2 Vols. Frankfurt/M: Suhrkamp.

Preiser-Kapeller, J. (2010a). Calculating Byzantium? Social Network Analysis and Complexity Sciences as tools for the exploration of medieval social dynamics. *Working Paper 'Historical Dynamics of Byzantium 1 (July 2010).'* http://www.oeaw.ac.at/byzanz/historical-dynamics.htm (accessed September 2010).

Preiser-Kapeller, J. (2010b). Calculating the Synod? A network analysis of the synod and the episcopacy in the Register of the Patriarchate of Constantinople in the years 1379–1390. *Working Article 'Historical Dynamics of Byzantium 2 (August 2010).'* http://www.oeaw.ac.at/byzanz/historicaldynamics.htm (accessed 1 September 2010).

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press.

Wetherell, C. (1999). Historical Social Network Analysis. In Griffin, L. J. and van der Linden, M. (eds), *New Methods for Social History*. Cambridge: Cambridge University Press, pp. 125–44.

Notes

1 Wetherell (1999) explains why 'historians . . . have been slow to adopt the [social network analysis] approach' and gives an overview of problems and prospects. Gould (2003) too mentions the 'mercifully slender portfolio of network analytic historical research' and gives examples of successful contributions of this method. For a brilliant use of SocNet analysis and visualization of its results, see the recent research by Preiser-Kappeler (2010a,b).

2 In social theory, communication is considered to be the basic operation of social systems. Luhmann (1997) describes society as 'ein auf Basis von Kommunikation operativ geschlossenes Sozialsystem'. Habermas (1988) too develops his theories based on a theory of communication.

3 The ALiVE visualization lab at the University of Hongkong is using our dataset to produce 3D visualizations of the social networks in immersive environments. See Kenderdine (2011).
4 The *Biographies of Eminent Nuns* [Biqiuni Zhuan 比丘尼傳 (T. 2063, dated 516)] has been added as of January 2011, and we are working on an eighteenth century collection of lay-persons-compiled by Peng Shaosheng 彭紹升 (1740–1796): *Biographies of Buddhist Laymen* (Jushi Zhuan 居士傳, CBETA/X.1646), the remnants of the earliest Buddhist biographical collection *Excerpts from the Biographies of Famous Monks* (Mingseng Zhuan Chao 名僧傳抄 CBETA/X. 1523), and the long and largely unexplored *Additions to the Continued Biographies of Eminent Monks* (Bu Xu Gaoseng Zhuan 補續高僧傳 CBETA/X.1524).