Interactive, Browser-based Information Foraging in Heterogeneous Space-Centric Networks

Alexander Savelyev[1,2] and Alan M. MacEachren[1,3]

Abstract – Social media data can be thought of as a multi-dimensional dataset that contains references (explicit and implicit) to a myriad of entities of different types: events, places, people, etc. Although there is a wealth of ideas about how to visualize individual dimensions of this heavily interconnected dataset (e.g. a variety of cartographic techniques can be used to explore the spatial properties of the data), attempts to explore the links among these dimensions face a number of both theoretical and practical challenges. A recent development in network science – heterogeneous network modelling – presents a number of unique solutions to some of those issues. This paper explores the potential and challenges associated with integration of the heterogeneous network modelling approach into interactive, visual analytical environments that focus on geographic data. We present a functional prototype implemented as part of SensePlace2, a web-based social media geovisual analytics environment.

Index Terms – geovisual analytics, heterogeneous network, information foraging, web-based.

INTRODUCTION

Social media data can be thought of as a multi-dimensional dataset that contains references (explicit and implicit) to a myriad of entities of different types: events, places, people, etc. Although there is a wealth of ideas about how to visualize individual dimensions of this heavily interconnected dataset (e.g. a variety of cartographic techniques can be used to explore the spatial properties of the data), attempts to explore the links *among* these dimensions face a number of both theoretical and practical challenges. A recent development in network science – *heterogeneous network modelling* – presents a number of unique solutions to some of those issues. This paper explores the potential and challenges associated with integration of the heterogeneous network modelling approach into interactive, visual analytical environments that focus on geographic data. We present a functional prototype implemented as part of *SensePlace2*, a web-based social media geovisual analytics environment.

1 HETEROGENEOUS NETWORK MODELING

A network – a group of entities connected to each other – is an abstract structure that can be used to model a wide range of phenomena. Some of the typical examples include social interactions (professional networks, personal contacts, Social Media relationships, etc.), transport infrastructure (including physical and information transit networks) and the like. Networks are composed of *nodes* and *links*, with nodes typically thought of as a representation of specific entities or concepts, and links used to establish connections between those entities. If we take the phenomenon of research collaboration as an example, the nodes in the research collaboration network would correspond to authors, publications, venues, conferences, etc. The links in this network would correspond to relationships between these nodes, such as "collaborates with" for pairs of authors, "attends" for authors and conferences, etc.

We can assume two different philosophies when we set out to analyse our research collaboration network. According to the first philosophy, we assume that all nodes and links are made equal, that is, our network is *homogenous*. In our particular example, this approach makes little sense – the semantics of the relationship between an author and a paper is quite different from that of an author and a conference. Although a homogenous network model does not fit our phenomenon of choice, this approach can be quite successful in simpler networks and is used in most of the current research on network science [1].

According to the second philosophy, nodes and links come in different types, that is, our network is *heterogeneous*. Now, what is the practical implication of this approach? It is clearly a more natural fit for our phenomenon (e.g. we can discern between "author" nodes and "paper" nodes and display them accordingly), but the key advantage is in using the additional semantic information to refine the network analysis algorithms. Every network analysis technique (e.g. ranking, clustering, similarity analysis, etc.) can be improved by the increased semantic resolution of network relationships that a heterogeneous network model provides [2, 3, 4].

There are two central concepts to heterogeneous networks – *network schema* and *meta-path* [1, 5, 6]. Network schema describes the connections that are possible in a given network. In our example, authors can generate papers and attend conferences, papers can be submitted to publication venues, etc. The network schema does not describe which particular author attended a specific conference, it only highlights the possibility of such relationship. Meta-paths describe the series of steps that can be taken across the network schema (e.g. author – paper – venue) in order to establish complex, composite relationships between different node types. Consider the following meta-paths based in our collaboration network example: "*author – paper – author*" and "*author – paper – venue – paper – author*". Both meta-paths establish a connection between pairs of authors, but the semantics of each connection are quite different – co-authors versus authors with papers at the same event.

Despite significant progress in the development of network analysis and data mining techniques associated with using heterogeneous network models, two areas of potential improvement can be identified. First, the semantic complexity associated with analyses based on meta-paths and network schema make it a logical candidate to be used in visual analytics environments that take advantage of human reasoning capacity for guiding the computational process [7].

Second, the applicability of a heterogeneous network philosophy to geographic data analysis is not explored, despite the fact that modern-day networked datasets are increasingly spatial in nature.

The following section describes our efforts in both of these areas, as we explore the applicability of the heterogeneous network

1. *GeoVISTA Center, Department of Geography, The Pennsylvania State University.*
2. *E-mail: savelyev@psu.edu.*
3. *E-mail: maceachren@psu.edu.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014.
For information on obtaining reprints of this article, please send *e-mail to: tvcg@computer.org.*

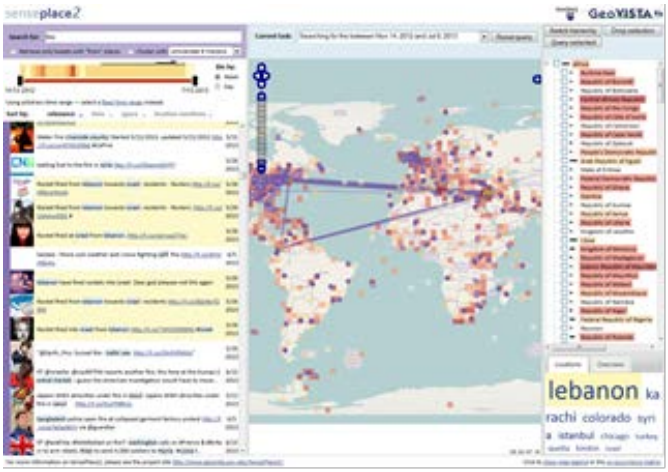

Fig. 1. Overview of SensePlace2 user interface (UI).

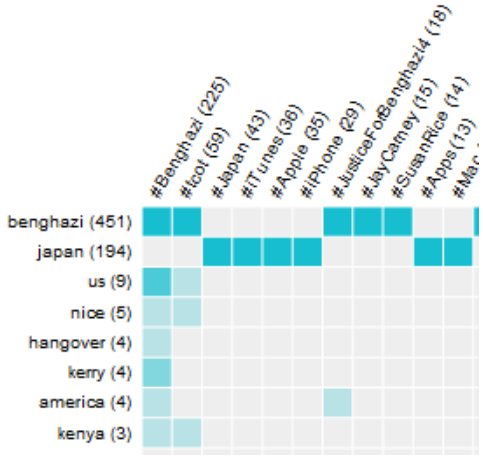

Fig. 2. Snippet of the co-occurrence matrix visualization.

modelling approach to interactive information foraging in space-centric networks.

2 Interactive Information Foraging in Space-Centric Networks

SensePlace2 is a web-based social media geovisual analytics environment that aims to improve situational awareness in crisis management and related application domains and that uses Twitter as the main source of geospatial information [8]. SensePlace2 makes use of both native Twitter metadata (such as message timestamp, GPS coordinates, hashtags mentions, re-tweets and @mentions) and Named Entity Recognition (NER) tools that extract references to places, people, organizations, etc. Place references are additionally processed using GeoNames-based geocoding tool and are assigned a place in the GeoNames place hierarchy. As a result, SensePlace2 deals with a high-dimensional dataset that consists of implicit and explicit place references at multiple scales together with temporal phenomena and thematic phenomena, such as trends in conversation topics and hashtag mentions.

SensePlace2 has a number of tools designed specifically to explore and analyse these different dimensions. Some of these tools are shown in Figure 1 above, namely a word cloud, a GeoNames hierarchy tree, a map with custom symbology, a timeline and a list of tweets relevant to the latest query (details of SensePlace2 implementation can be found in the paper cited above). One of the tools introduced recently in order to explore the links among these dimensions is a co-occurrence matrix, shown in Figure 2 above.

The purpose of the co-occurrence matrix is to explore the relationship between any given pair of entity types found in SensePlace2. The screenshot above displays the correspondence between place mentions and hashtag use, as found in the results of the latest query. In this particular dataset, for example, a number of hashtags are strongly associated with references to Benghazi and Japan. The co-occurrence matrix is the first tool of the SensePlace2 suite to use a heterogeneous network as its data model, as explained below.

2.1 Heterogeneous Network Modelling in SensePlace2

2.1.1 Heterogeneous Network Schema

As mentioned in Section 1 above, one of the key aspects of the heterogeneous network is its *schema*. A network schema describes the connections that are possible in a given network, and Figure 3 (seen on the next page) demonstrates the network schema used in SensePlace2. As seen in Figure 3, the network structure is centred around the notion of a *tweet*, sent by a particular *user* (the tweet's author), as well as a range of metadata associated with that tweet: the *hashtags* that were included as part of the message, the *GPS coordinates* associated with the tweet (if any), the *place names* mentioned in that particular tweet, the *continent* and the *country* to which the place mention has been assigned as part of the geocoding process, etc.

As explained earlier, a network schema describes the connections that are possible within a given network, but does not describe the connections of any entity in particular. An example of what an actual network would look like is shown in blue in Figure 4 (seen on the next page as well). For the sake of brevity, Figure 4 illustrates a small subset of an actual network and only has three types of nodes in it – place mentions, tweets and users.

A structure similar to that shown in Figure 4 is used by the co-occurrence matrix to calculate the strength of relationship between different entity types. Using our miniature network, we can show how co-occurrence between place mentions can be established, that is, we can determine *which places tend to be mentioned together*. Using the miniature network shown in Figure 4 below, what places is *Ghana* mentioned together with? The most obvious answer is "US and London", as these three locations were mentioned in the same tweet. However, there are other answers possible. Instead of asking a fairly narrow question "what places are mentioned whenever Ghana is referred to in a tweet", we can ask a broader question such as "what places are referred to by people who mentioned Ghana at least in one of their tweets". In this case, the answer will be "US, London, Paris and UK", as Bob, the author of the tweet about Ghana, mentioned all of them in his earlier tweets. The difference between these two questions can be formally described using the second key concept of the heterogeneous network modelling process – the *meta-path*.

2.1.2 Meta-Paths

Meta-path can be thought of as the prescribed set of directions for getting across the network schema from one entity type to another. In our miniature network example, we are interested in establishing connections between place mentions. Using Figure 4, we can define the following ways of doing so: *place mention – tweet – place mention*, or *place mention – tweet – user – tweet – place mention*, or even *place mention – tweet – user – user – tweet – place mention*. These directions are the three possible meta-paths that exist between place mentions in our schema (excluding paths with loops). Notice that the three paths defined above have different semantics. The first path can be interpreted as "co-occurrence of place mentions within a single tweet" (Figure 6), the second path can be interpreted as "co-occurrence of place mentions across all tweets by the same user" (Figure 7), and the third path can be interpreted as "co-occurrence of place mentions across all tweets by users within an established user community".

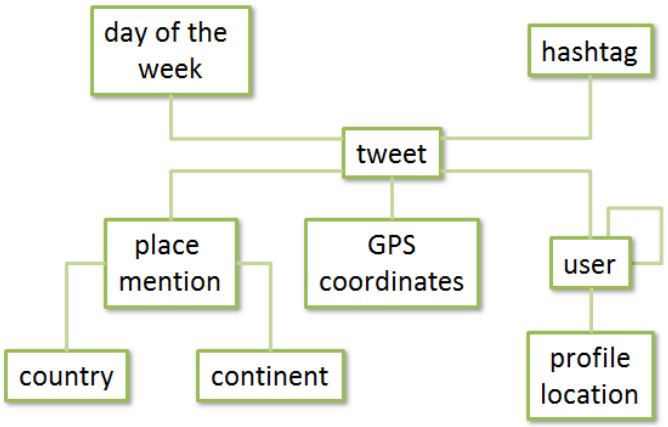

Fig. 3. SensePlace2 heterogeneous network schema.

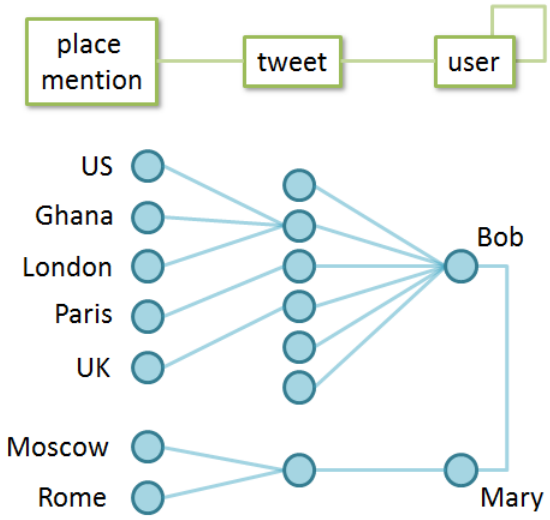

Fig. 4. Sample heterogeneous network.

Fig. 5. Meta-path selection tool.

The logic outlined above applies equally well to any other combination of entity types found in the SensePlace2 network schema. For example, we can define meta-paths to explore the strength of connection between hashtags within a single tweet, hashtags across all tweets by the same user, or hashtags mentioned across all tweets within existing user communities. Looking at place mentions once more, we can explore the strength of connections between place mentions that are accompanied by the same hashtag (*place mention – tweet – hashtag – tweet – place mention*), are brought up on the same day of the week (*place mention – tweet – day of the week – tweet – place mention*), or tend to be brought up by people from the same area (*place mention – tweet – user – profile location – user – tweet – place mention*). Meta-paths need not be symmetric – the screenshot shown in Figure 2 was generated using a meta-path *place mention – tweet – hashtag*.

2.2 Integration of Heterogeneous Network Model in Geovisual Analytics Environment

A prototype of the heterogeneous network foraging system has been implemented as one of the SensePlace2 analytical components. The network itself is built in-memory using JavaScript code and contains the entire network of connections between entities found in the latest query results (1000 most relevant matches to the query), with an average size of approximately 6,000 nodes per 1,000 tweets. Two important components of this prototype that were not described in earlier sections are the *meta-path selection tool* and the integration of heterogeneous network model with the rest of the SensePlace2 UI.

2.2.1 Meta-Path Selection Tool

One of the key advantages of a heterogeneous network modelling approach is that meta-paths, its primary query tool, can be easily defined by the analyst. A prototype meta-path selection tool is demonstrated in Figure 5 above. The network path currently selected is *place mention – tweet – user – tweet – place mention* (its semantics have been explained above). The meta-path selection tool works by offering the analyst a blank drop-box menu that contains the list of all entity types in the network schema. As the analyst makes their choice (with the *place mention*, labelled as *"about" location* in the illustration above, being the first pick), another drop-box is added to the meta-path string. This time, the analyst will be limited to only choosing between the entity types that are connected to the *place mention* (their first pick) in the network schema. By making their pick in the blank drop-down boxes, the analyst is essentially "stepping" along the links defined in network schema. The co-occurrence matrix visualization (shown in Figure 2 above and Figures 6 and 7 below) is dynamically redrawn every time a choice is made, showing the strength of connection between the first and the last item in the selected meta-path (the colour of the cells is represents the number of connections between the specific pair of entities). In the example shown above, the analyst made 5 choices so far to explore the strength of connection between nodes of type *place mention*. The analyst is offered another drop-down menu in case they intend to continue building the meta-path.

As a shortcut, the analyst can also make a pick of row and column entities for the co-occurrence matrix right away without bothering with the meta-path selection (the two drop-down boxes at the top of Figure 5). In this case, the meta-path is ignored and the connection between the two entity types is established using the shortest path on the network schema.

2.2.2 Integration with SensePlace2 UI

The heterogeneous network foraging system described in the previous section is a powerful tool, but it becomes most useful when used in combination with other components of the SensePlace2 UI, such as its word cloud component, its map and the tweet list. Because of its screen space requirements, the co-occurrence matrix is implemented in as a browser pop-up window that can be positioned alongside the main application, and is fully linked to it by means of the SensePlace2 UI *coordination framework*.

The SensePlace2 coordination framework is a multiview user interface coordination mechanism implemented in JavaScript specifically for interactive browser-based environments [9]. The key principle of its operation (technical details can be found in the paper cited above) is that SensePlace2 components are built to adhere to a

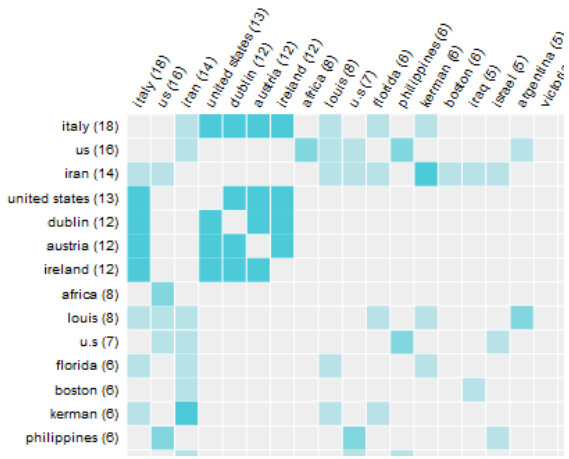

Fig. 6. Places mentioned within the same tweet.

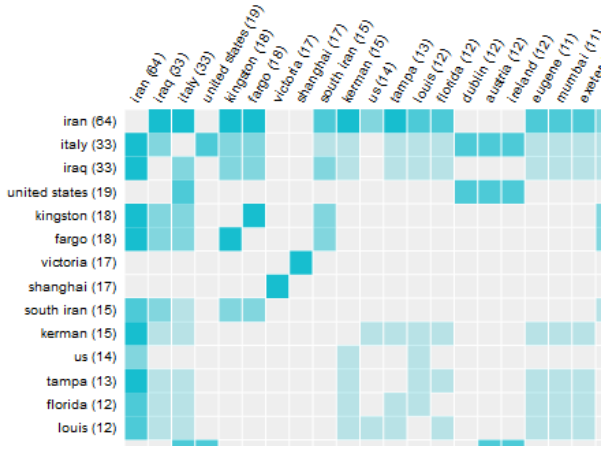

Fig. 7. Places mentioned by the same user.

standardized coordination protocol and are linked together by a dedicated coordination manager. The final outcome of this process is that any user interaction with one of the SensePlace2 components is automatically propagated through the rest of the interface.

For example, if the analyst was to click on the co-occurrence matrix cell that is at the intersection of *Japan* and *#Apple* entities in Figure 2, the rest of the interface will be re-drawn to reflect this selection: the map will show the locations of tweets that mention Japan and contain the hashtag #Apple, the tag cloud will show the terms frequently mentioned alongside references to Japan and #Apple, etc. The coordination mechanism is symmetric, which means it is possible to click on the term "Lebanon", as shown in Figure 1, and the co-occurrence matrix will be adjusted to only show connections between entities that occur in tweets mentioning Lebanon.

The heterogeneous network foraging system described in this section becomes a potent multi-dimensional exploration tool when coupled with interactive mouse-over and filter capacity of the SensePlace2 coordination mechanism; it follows the "overview – filter – detail on demand" metaphor well.

3 Limitations and Future Directions

Current implementation of the heterogeneous network foraging system described above is a prototype built to explore the potential and challenges associated with integration of the heterogeneous network modelling approach into interactive, visual analytical environments that focus on geographic data. Although it fulfils its role quite well, this prototype is limited in a number of important ways that will be addressed as part of future research.

The most important limitation is that we currently use the query results (limited to about 1,000 tweets) to build the heterogeneous networks presented to the analyst. Building such a network using our entire multi-year data archive requires significant investments into graph database technology and accompanying hardware, which could be unwise before the potential utility of the resulting foraging system is fully evaluated.

This limitation is somewhat offset by the fact that the network foraging system implementation is built using the Linked Data principles with every entity in the SensePlace2 system mapped to a unique Uniform Resource Identifier (URI), which makes detail-on-demand queries possible even with limited investment in hardware.

Another important limitation of this study is the lack of an empirical user evaluation of the meta-path selection tool and the associated workflow. This study will be performed as part of continuing research on this project.

Acknowledgments

This material is based upon work supported by the US Army Engineer Research and Development Center (ERDC), Topographic Engineering Center (TEC) under Contract No. (Contracting agency(ies) contract number) W912HZ-12-P-0334.

References

[1] Sun, Y., & Han, J. 2012: Mining Heterogeneous Information Networks: Principles and Methodologies. Synthesis Lectures on Data Mining and Knowledge Discovery, 3(2), 1–159.

[2] Gahegan, M., Agrawal, R., Banchuen, B. and DiBiase, D. 2007: Building rich, semantic descriptions of learning activities to facilitate reuse in digital libraries. International Journal on Digital Libraries 7, 81-97.

[3] Gahegan, M., Agrawal, R., Jaiswal, A., Luo, J. and Soon, K. 2008: A Platform for Visualizing and Experimenting with Measures of Semantic Similarity in Ontologies and Concept Maps. Transactions in GIS 12, 713-732.

[4] Gahegan, M., Luo, J., Weaver, S.D., Pike, W. and Banchuen, T. 2009: Connecting GEON: Making sense of the myriad resources, researchers and concepts that comprise a geoscience cyberinfrastructure. Computers & Geosciences 35, 836-854.

[5] Sun, Y., & Han, J. 2013: Meta-path-based search and mining in heterogeneous information networks. Tsinghua Science and Technology, 18(4).

[6] Shen, W., Han, J., & Wang, J. 2014: A Probabilistic Model for Linking Named Entities in Web Text with Heterogeneous Information Networks.

[7] Weaver, C. 2010: Cross-filtered views for multidimensional visual analysis. Visualization and Computer Graphics, IEEE Transactions on, 16(2), 192–204.

[8] MacEachren, A. M., Jaiswal, A., Robinson, A. C. Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., and Blanford, J. 2011: Senseplace2: Geotwitter analytics support for situational awareness. Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on, 181-190. DOI= http://dx.doi.org/10.1109/VAST.2011.6102456.

[9] Savelyev, A. 2013: Multiview user interface coordination in browser-based geovisualization environments. Proceedings of the 1st ACM SIGSPATIAL International Workshop on MapInteraction (pp. 54–58). ACM.