Visualizing Overlapping Latent Communities Using POI-Based Visualizations

Patrick M. Dudas
University of Pittsburgh
pmd18@pitt.edu

Jae-wook Ahn
Drexel University
ja626@drexel.edu

Martijn de Jongh
University of Pittsburgh
m.a.dejongh@gmail.com

Peter Brusilovsky
University of Pittsburgh
peterb@pitt.edu

Abstract

Social network analysis and social network visualizations can provide a meaningful statistical and topological understanding of latent communities. However, the majority of current visualization approaches just represent sub-communities as clusters of closely related nodes in a node-link diagram and embed limitations to represent overlapping communities and multi-layer community structure frequently found from modern complex networks. We argue that visualizations based on points of interest can provide a better solution to represent overlapping latent sub-communities. We present two visualization systems, SuperVIBE and ContextForces, which implement this approach. These systems operate by creating two-dimensional latent spaces by means of grouping nodes using external variables not presented in the graph and by offering an interactive visualization to filter and map in these latent spaces. Understanding which latent groups are most central to a variety of topics and providing visual clues to the individuals critical to those groups provides a mechanism to explore and discover overlapping latent communities.

Keywords : social network analysis, latent community, visualization, SuperVIBE, ContextForces

Introduction

In recent years, social network analysis and visualization have emerged into a very popular topic (Fortunato, 2010). A reasonable fraction of research on this topic focuses on visualizing a set of latent (sub)communities within the rich structure formed by social connections among people. It is interesting, however, that the last 10 years brought almost no changes to the dominant approach of presenting the structure of discovered communities: the majority of modern work still present sub-communities as clusters of closely located nodes in a traditional node-link diagram (Newman & Girvan, 2004; Wakita & Tsurumi, 2007). The only major change – a gradual move from static to interactive (exploratory) social network analysis (SNA) that offered users an ability to interact and manipulate the graph or statistical variables to highlight key nodes or leaders (Perer & Shneiderman, 2008) – has not augmented or improved the visualization approach itself.

Social networks explored by modern researchers are typically much more complex than simple node-link diagrams. These "modern" social networks are based on larger and more heterogeneous data (frequently extracted from online systems) and frequently represent several different dimensions of similarity among people. To deal with this increased complexity and heterogeneity, data mining researchers introduced more and more sophisticated community-mining approaches. A number of these approaches were specifically created to discover overlapping communities and multi-layer community structure. In this context, the dominated node-link visualization approach created for presenting simpler non-overlapping communities emerged as a bottleneck, restricting our ability to analyze and comprehend modern complex community structures.

Dudas, P. M., Ahn, J-w., de Jongh, M., & Brusilovsky, P. (2013). Visualizing overlapping latent communities using POI-based visualizations. *iConference 2013 Proceedings* (pp. 601-610). doi:10.9776/13287

In this paper we suggest two alternative approaches to visualizing complex, overlapping or multi-layered community structures based on interactive exploration with Points of Interest (POI). The POI-based visualization was originally created to visualize search results; however, we believe that with some moderate modification, this technology can provide a great approach for exploring (overlapping) latent communities. The paper presents two POI-based approaches for exploring latent communities: (1) SuperVIBE introduces a few changes to the original POI-based visualization known as VIBE; and (2) ContextForces attempts to bridge the gap between traditional node-link diagrams and a POI based visualization. The following section describes, in detail, the two visualizations, and concluding section looks at the future work and development of both of these projects.

Dataset

Network

To explore multi-layer social network visualizations, we used a dataset containing several different types of connections among a group of authors that have published in the UMAP (User Modeling, Adaptation and Personalization) conference series. The UMAP is the main conference series on adaptive systems. To create this dataset, we have extracted connection data from the DBLP database (Ley, 2002) and the Conference Navigator 3 (CN3) system. CN3 was developed at the University of Pittsburgh's iSchool (Parra, Jeng, Brusilovsky, López, & Sahebi, 2011) and it supports conference organizers and delegates with web-based informational, scheduling, and social features. Social features let the users see each other’s schedules and connect using CN3 networking options (make a bidirectional connection or follow somebody). It also provides third party social network information, including Facebook, LinkedIn, CiteULike, BibSonomy, and Mendeley. CN3 has been used at several UMAP conferences and we extracted co-bookmarking data to socially connect authors, if they bookmarked similar sets of papers within the CN3 system. We also used the author’s publication data in order to connect them if they published articles with similar topics. The social networks were created from the data by aggregating co-authorship information and similarity measures into links, and storing authors as nodes.

Latent Communities as POIs

The main challenge of any POI-based approach is the selection of meaningful POIs. While there are many possible options for POI selection in the context of multi-layer communities, we started our exploration of multi-layer community visualization with a specific POI selection approach that we consider as most promising and universal. This approach is based on two principal ideas: (1) the POIs used for visualization correspond to the top 10 or 20 latent communities extracted from a multi-layer social network; (2) the latent communities (i.e., POIs) are discovered using Latent Dirichlet Allocation (LDA) algorithm (Blei & Lafferty, 2007). LDA is an algorithm widely used for probabilistically discovering topics from a set of documents based on keyword frequencies.

Following this approach, we applied LDA to the multi-layer data formed by UMAP authors’ publication or bookmarking activities in order to elicit latent communities represented as research topics. The process is as follows:

(1) Extract papers the authors bookmarked or published;
(2) Convert the papers to a bag-of-word representation (author-keyword matrix);
(3) Feed the matrix to LDA;
(4) Two datasets (bookmarking and publication) are created using LDA. Each includes two outputs: (a) topic-probability pairs per each author; (b) list of keywords with higher probabilities per each topic.

Output (a) maps the authors and the topics (latent communities represented as POIs) and (b) provides the keyword of each community (e.g. “social, search, web” in Figure 1).

POI-Based Overlapping Latent Community Visualizations

The social network community discovery methods mostly rely on graph structures by dividing network nodes into densely connected subgroups or clusters (Newman & Girvan, 2004; Wakita & Tsurumi, 2007). They are more appropriate for representing one community for each person in the network and cannot support overlapping latent communities, where community membership is not decided by explicit conditions (*latent*) and a single person can participate in multiple communities (*overlapping membership*). An example of these types of communities is an academic conference, where people gather and form latent communities following their research interests but these communities may not be consistent with their physical affiliations. Rather, their activities such as co-bookmarking behaviors within a conference management system can better aid such community discovery task. The membership derived from these activities is not explicitly defined and one person can be a member of multiple communities. Our SuperVIBE and ContextForces systems are devised to best represent this type of community. They are expected to meet the following requirements:

1. Visualize multiple latent community membership of a single person.
2. Visualize the different degree of membership. One person can be more strongly related to one community than the others.

Latent communities represented as POIs SuperVIBE and ContextForces can fulfill these requirements. People are placed closer or further from the multiple POIs (Requirement 1), according to their degree of membership (Requirement 2). Additionally, ContextForces keeps intact the node-link paradigm to add an additional layer of network connectivity. The details are introduced below.

SuperVIBE

VIBE (Visual Information Browsing Environment) is a POI-based visualization developed by (Olsen, Korfhage, Sochats, Spring, & Williams, 1993). It displays the POIs and related objects by locating the objects according to the similarity ratio to the POIs. An object that is more similar to a POI is placed closer to the POI and the distances to POIs are always consistent with the similarity ratios. VIBE can visualize N-to-N relationships (objects-to-POIs) and it makes VIBE suitable for visualizing overlapping people-to-latent communities relationships. We expect VIBE can benefit the following tasks:

1. Discover and rank the top N members for overlapping latent communities.
2. Discover which communities are more relevant for a member.

Figure 1 shows the relationships between the latent communities (POIs in yellow circles) and the authors (squares) in VIBE. The POIs are labeled with latent community concepts discovered by LDA using the authors' co-bookmarking actions. Despite the straightforwardness of the VIBE algorithm, its one shortcoming is that it is based on similarity ratios and easily creates clutter when the differences between the similarity value ratios are smaller (Figure 1). Therefore, we extended the traditional VIBE as SuperVIBE by modifying the visualization algorithm as follows:

1. Consider only the POIs that have the similarity values greater than a specified threshold.
2. Consider only the top N POIs for a user.

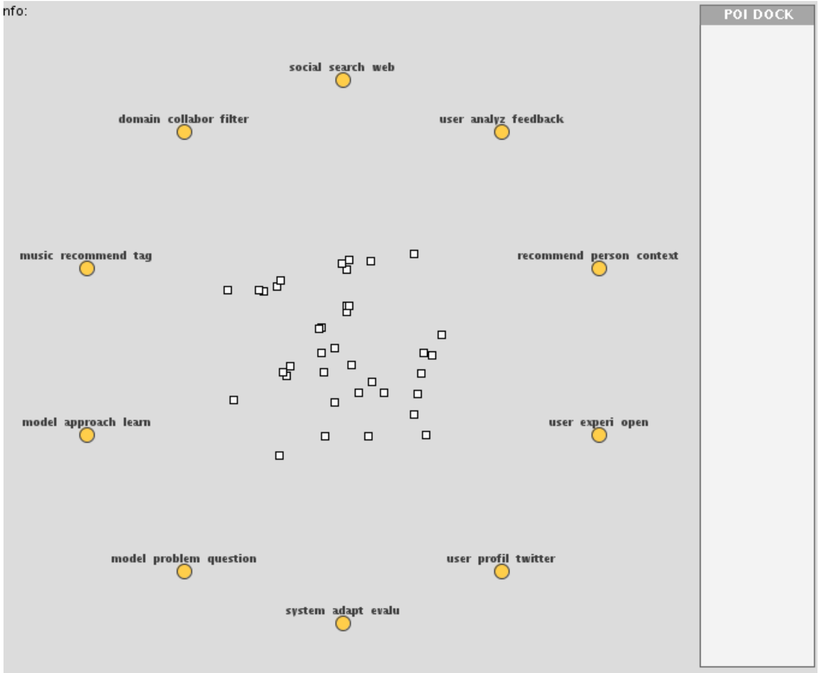
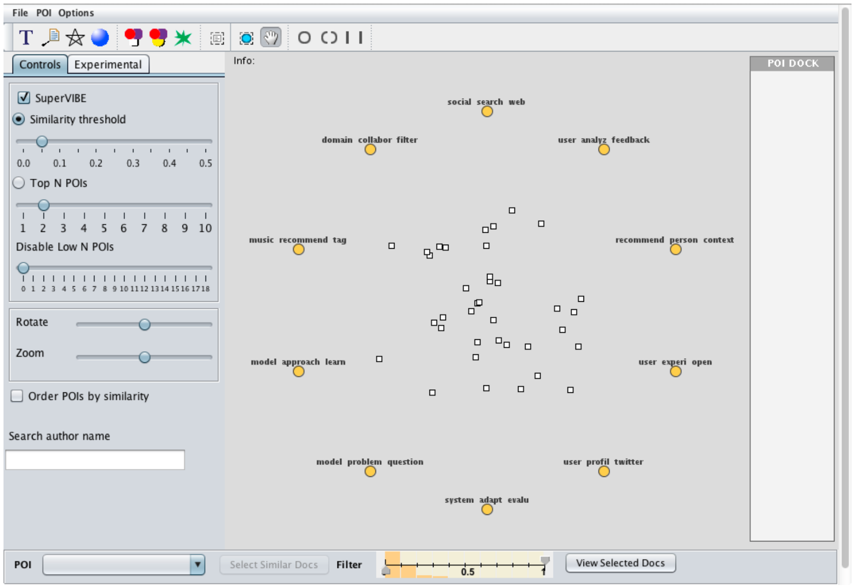


Figure 1. VIBE visualization of co-bookmarking communities. POIs (circles) are communities labeled with key-concepts and squares are authors. Most are cluttered in the center.

These features reduce the influences of the POIs with lower membership similarity and make the users move towards the ones with higher similarities. We allow SuperVIBE users to switch between the two features, in order to discover the best visualization ("Controls" in Figure 2 & 3). Figure 2 shows that it can reduce the clutter in the center by considering the POIs that have similarities greater than a threshold (0.05).

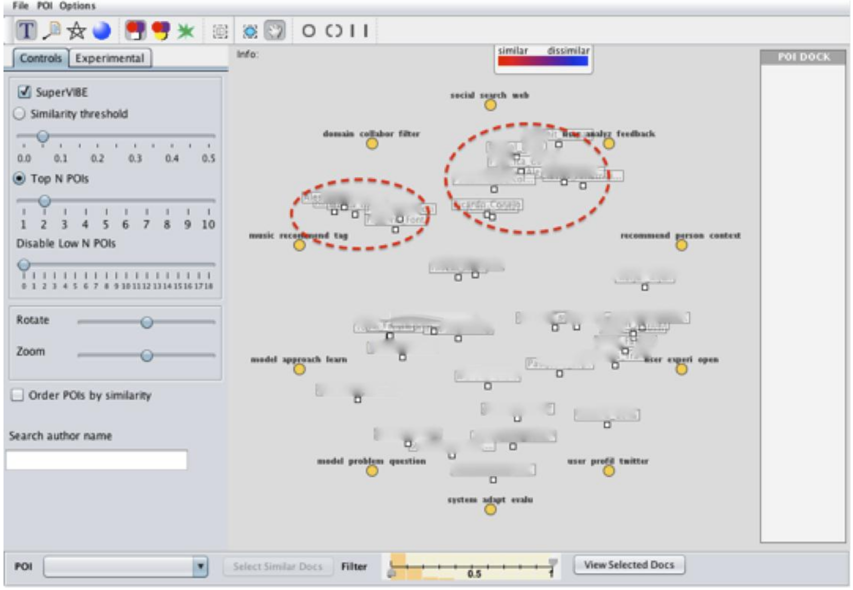


☑ SuperVIBE
● Similarity threshold
○ Top N POIs
Disable Low N POIs
□ Order POIs by similarity

Figure 2. SuperVIBE using the "similarity threshold" method

Figure 3 shows a more revealing visualization. This time, it considers the top N POIs from each author when calculating their positions. It clearly shows several clusters gathered around the POIs. One big cluster is found just below the "SOCIAL SEARCH WEB" and "USER ANALYZE FEEDBACK" latent community POIs, which suggests that the authors with the interests in those topics belong to the overlapping latent communities. Another example is a cluster of five people that are gathered around

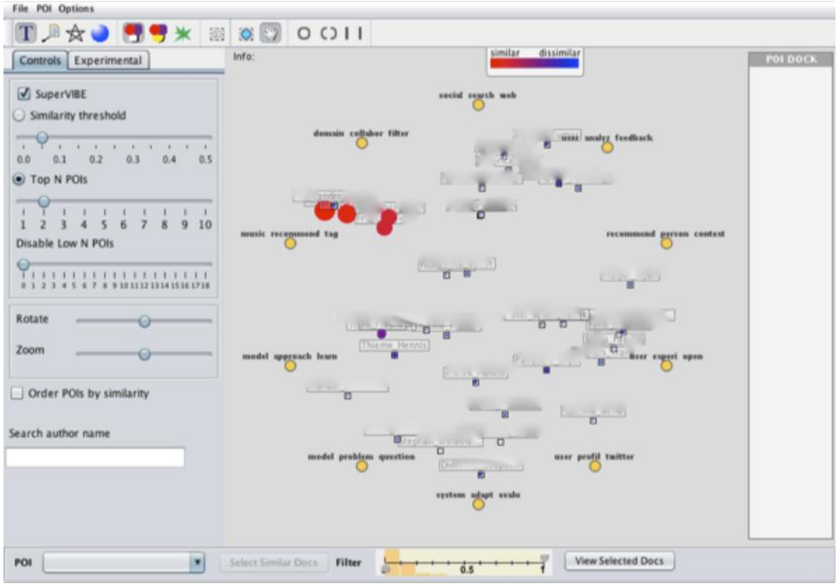
“MUSIC RECOMMEND TAG” and “DOMAIN COLLABOR FILTER.” We can assume that these authors are involved in music recommendation and collaborative filtering communities.



☑ SuperVIBE
○ Similarity threshold
● Top N POIs
□ Order POIs by similarity

Figure 3. SuperVIBE Using Top N POI Method

It also supports visually discovering community members from POIs. Figure 4 shows an example of searching for people who might belong to the “MUSIC RECOMMEND TAG” community. When the mouse cursor is moved over a POI, the authors that are highly similar to the POI are highlighted. The size of the highlights (bigger, more similar) and the color (red, more similar) corresponds to the strength of a similarity. Therefore, we can see that the 4 people with big, red highlights are the most relevant members of the community. Even though the proximity in the visualization (POI to authors) can also support the task, this color-based method can help users complete the task more quickly.



☑ SuperVIBE
○ Similarity threshold
● Top N POIs
□ Order POIs by similarity

Figure 4. Finding members from a community POI

ContextForces

ContextForces provides both the mechanism of POI-based visualization and traditional node-link diagram using force directed placement (FDP). It was designed using an open-source project called

D3.js (Bostock, Ogievetsky, & Heer, 2011). For examining the roles of actors in a social network, a plethora of the research has focused primarily on the development of groups, communities, cliques, and mutually-relevant, homogeneous clusters of individuals based on metadata about the actors (Ahn, Han, Kwak, Moon, & Jeong, 2007; Heer & Boyd, 2005; Matsuo et al., 2006). However, this is mostly tapered to statistical calculations or user-generated interpretation. ContextForces' POI-based approach applies the properties of a Venn diagram (e.g. InfoCrystal (Spoerri, 1993)) to the baseline FDP. It relies on the FDP (Fruchterman & Reingold, 1991) framework to designate the placement of the nodes as the user interacts with the visualization. In a typical FDP environment, nodes are placed based on the edges and the weighting between each edge. For example, Figure 5 presents a synthesized FDP arrangement.

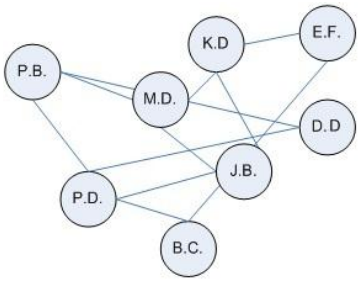


Figure 5. Synthesized FDP network

Figure 6 presents the same arrangement after POIs are applied and are locked to the apex of a circular shape, arranged in equal degrees from one another around the network. In this figure, nodes are augmented to reflect these variables and will render themselves to those new positions based on the weights placed by the POIs. The resulting position of the nodes is produced by taking into account both traditional FDP placement and POI attraction. To stress the impact of POIs, nodes can be viewed in a Venn-diagram like manner, highlighting the interaction of the external variables (Figure 7).

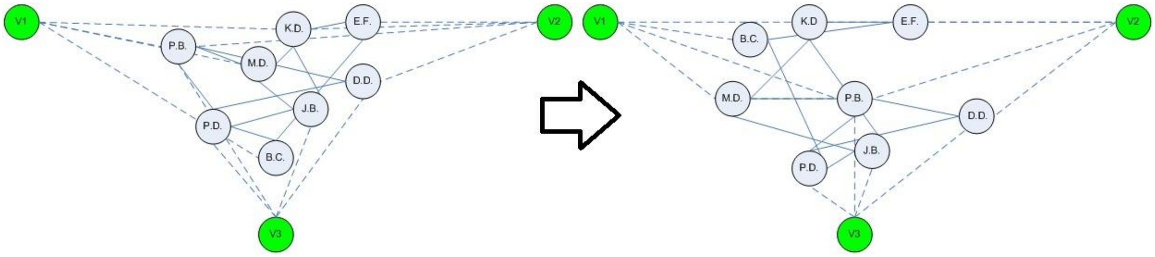


Figure 6. POIs pulling nodes from the FDP layout (based on weighted edges)

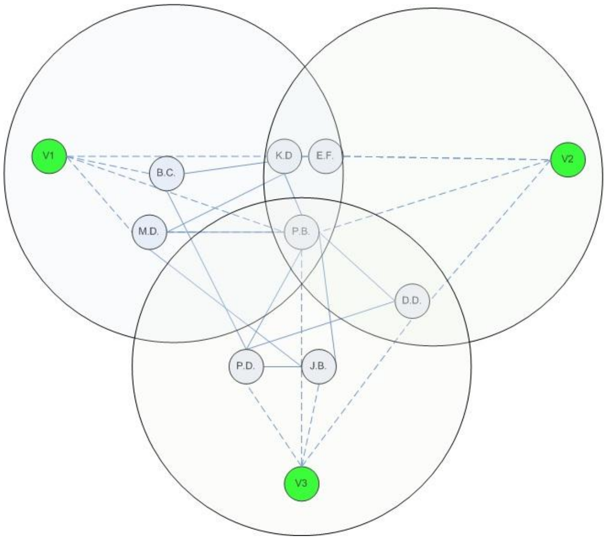


Figure 7. Venn diagram appearance using the synthesized FDP network

Researchers can apply any POIs to redefine the latent communities. ContextForces subdivides these very large networks (in the case of the UMAP dataset, 766 nodes and 19,000+ edges) into, at most, (n^2 – n + 1) subdivisions.

The concept of ContextForces is to add POIs and allow the visualization to be altered interactively so researchers can see how the POIs and the network are unified. At the same time, users can switch between the POI-based visualization and FDP approach and see how the network evolves. In this case, Affinity Propagation clustering algorithm (Frey & Dueck, 2007) is applied to the network and the subsequent color represent the community (or cluster) of each node. We implemented functionality to either relax or exacerbate either of these two mechanisms, including the:

1) Ability to shrink or elongate the distance between the nodes within a cluster or nodes attached to POIs (Figure 8)
2) Ability to increase the font size based on linearly increasing the value or based on degree centrality of each node in the original topology (Figure 9)

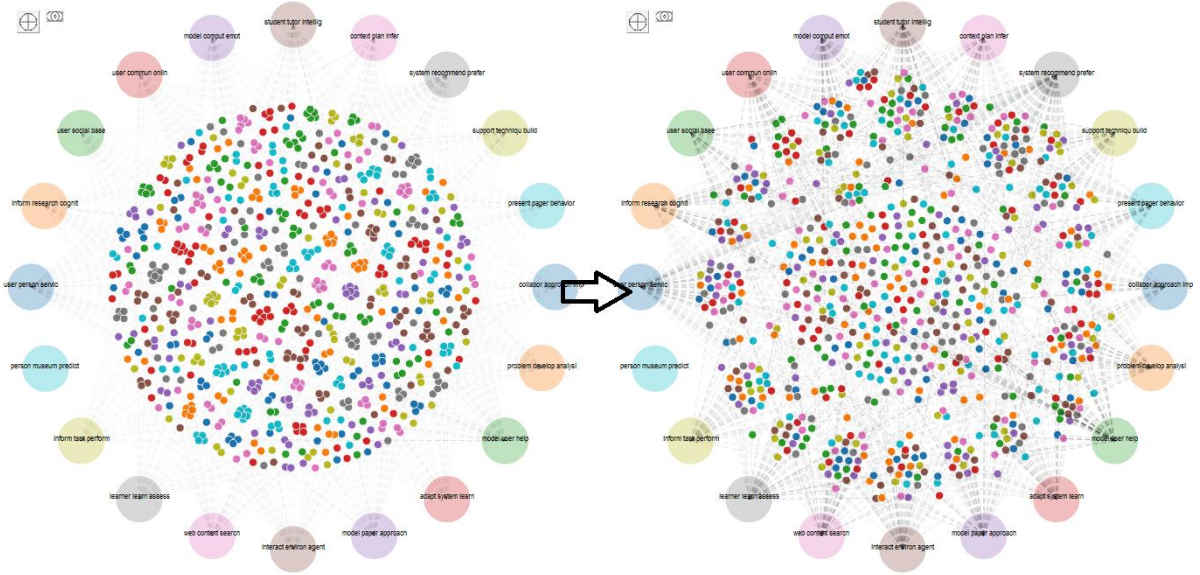


Figure 8. Shrinking links between inner-cluster nodes and POIs and nodes

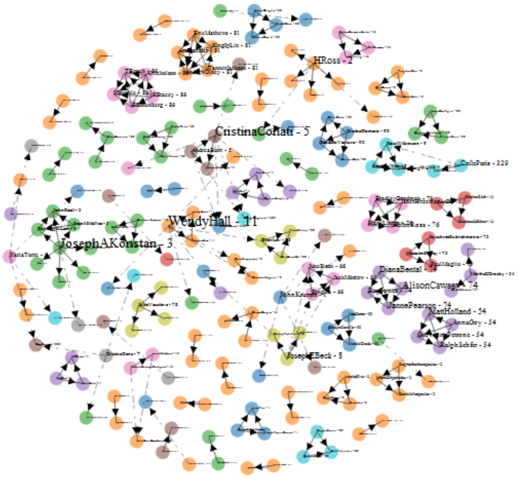


Figure 9. Degree centrality font

Lastly, we define edge weights for each author to each topic based on LDA values. We can then trim these edges to show better exemplar authors for each topic. Figure 10 shows the complete graph without trimming and Figure 11 shows the trimmed graph.

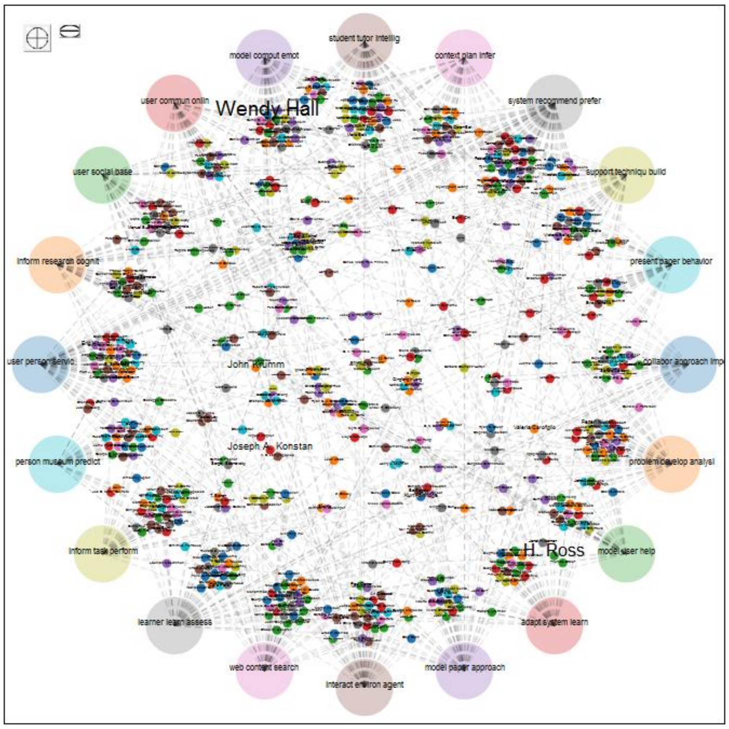


Figure 10. Complete graph and POIs

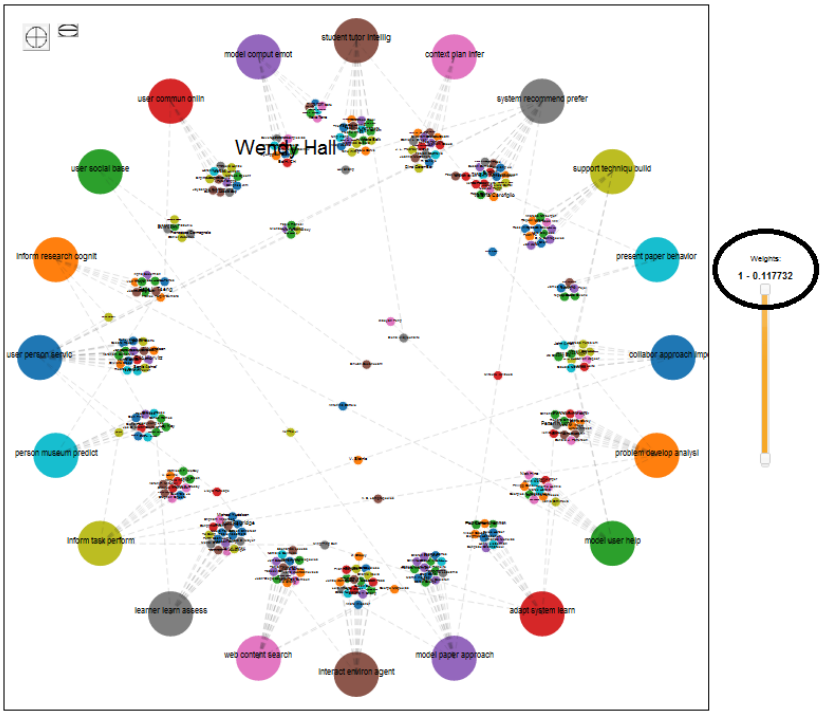


Figure 11. Trimmed graph based on weight

Conclusions

Discussed in the paper is a network based on the UMAP dataset and latent community structures determined by LDA. We propose two novel approaches to spatially visualizing the POIs representing these UMAP latent communities. SuperVIBE considers both the POIs that have the similarity values greater than a threshold and only the top N POIs for a user. ContextForces looks at these same two criteria, but also applies the node-link diagram using force directed placement. We then applied the UMAP dataset and the latent community structures determined by LDA to both visualizations to showcase a new paradigm in the visualization of social network analysis.

We expected that these two approaches could better define overlapping latent communities using spatial methods and promoting a new paradigm in mapping data with network visualization. We believe this mapped data is limitless and without bounds in terms of context in that any latent artifact can be mapped to each visualization and new insight can be found about network.

While we believe that the suggested POI-based approaches offers an important advantage to those exploring modern online communities, the true impact of this approach has to be determined in a user study. We will approach using comprehensive user-study to test whether our tools can achieve the expected goal and to identify which features allow for the identification of key individuals in a given network dataset in a timely fashion, among other network or overlapping latent community facets. We also plan to extract latent communities using additional social activities such as co-attending conferences. If this does warrant a valid hypothesis, we would then like to select other datasets that require identification of key players including: dark networks (i.e. suspected criminals or terrorists), medical data, conference authorship, and others. Utilizing these diverse datasets we will explore and develop novel visualization functions to further improve exploratory SNA.

References

Ahn, Y.Y., Han, S., Kwak, H., Moon, S., & Jeong, H. (2007). *Analysis of topological characteristics of huge online social networking services*.

Blei, D.M., & Lafferty, J.D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 17-35.

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D³ Data-Driven Documents. *Visualization and Computer Graphics, IEEE Transactions on, 17*(12), 2301-2309.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports, 486*(3), 75-174.

Frey, B.J., & Dueck, D. (2007). Clustering by passing messages between data points. *science, 315*(5814), 972-976.

Fruchterman, T.M.J., & Reingold, E.M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience, 21*(11), 1129-1164.

Heer, J., & Boyd, D. (2005). *Vizster: Visualizing online social networks*.

Ley, M. (2002). *The DBLP computer science bibliography: Evolution, research issues, perspectives*.

Matsuo, Y., Hamasaki, M., Nakamura, Y., Nishimura, T., Hasida, K., Takeda, H. (2006). *Spinning multiple social networks for semantic web*.

Newman, M.E.J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E, 69*(2), 026113.

Olsen, K.A., Korfhage, R.R., Sochats, K.M., Spring, M.B., & Williams, J.G. (1993). Visualization of a document collection: The VIBE system. *Information Processing & Management, 29*(1), 69-81.

Parra, D., Jeng, W., Brusilovsky, P., López, C., & Sahebi, S. (2011). Conference Navigator 3: An Online Social Conference Support System.

Perer, A., & Shneiderman, B. (2008). *Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis*.

Spoerri, A. (1993). *InfoCrystal: A visual tool for information retrieval & management*.

Wakita, K., & Tsurumi, T. (2007). Finding community structure in mega-scale social networks. *arXiv preprint cs/0702048*.