Automatic hypothesis generation and evaluation by network structure content analysis and visualization.

Don Pellegrino (Ph.D. Student), Chaomei Chen, Ph.D. (Co-PI, NEVAC) Drexel University
DHS COE: North-East Visualization & Analytics Center (NEVAC)

Introduction:

The Semantic Network Structure Analysis (SemanticNetSA) tool introduced here integrates visualization and content analysis techniques to uncover salient, novel and otherwise interesting patterns in network and graph data structures and applies those patterns to the discovery of similar themes in new data. SemanticNetSA uses Prefuse, SUBDUE and statistical analysis of GraphML data to present patterns of entity relations in context, thus helping analysts develop hypotheses that use network data as evidence. It is hypothesized that key events or entities will share similar structural relationships in the context of neighbor events or entities and that such patterns may be reused to propose hypotheses representing analogous themes in future data. For example the system may suggest a hypothesis of an emerging disease outbreak to an analyst by searching for patterns that have occurred in emergency room data during historical instances of outbreaks. The hypothesis will be constructed by its evidentiary network data, such as the emergency room data constituting the pattern and shown in context of correlating data to facilitate evaluation. The tool will help analysts explore the relationship between the development of hypotheses and the use of network data as evidence. Increased understanding in this area will lead to better automated support for hypotheses in analytical tools (Heuer 1999), (Thomas and Cook 2005).

Research questions:

1.Can objective interestingness measures identify patterns of relations that serve as useful initial hypotheses?
2.How should patterns of relations be captured to support reuse in hypothesis formulation?
3.Can visual navigation of graphs support evidence based evaluation of hypotheses?

Materials:

• Heterogeneous domain data transformed into associative network data.
•Visualization Technologies: Prefuse Visualization Toolkit
•Analytics Technologies: SUBDUE, Interestingness Measures

Results and conclusions:

Initial results are very encouraging. The prototype system successfully identifies substructures in graph data. The discovered structures have been found to be useful for developing hypotheses regarding the content of the data.

Literature cited:

Cook, D.J., & Holder, L. (2007). SUBDUE Manual Version 1.4 (pp. 40). Pullman, WA: Washington State University.

Heuer, R.J., Jr. (1999). Psychology of Intelligence Analysis: Central Intelligence Agency.

Thomas, J.J., & Cook, K.A. (Eds.). (2005). Illuminating the Path: The Research and Development Agenda for Visual Analytics. Los Alamitos, CA: IEEE Computer Society.

Acknowledgements:

This project was funded through the Regional Visualization and Analytics Centers (RVACs) by the Department of Homeland Security, Science and Technology Directorate, Office of University Programs.

For further information:

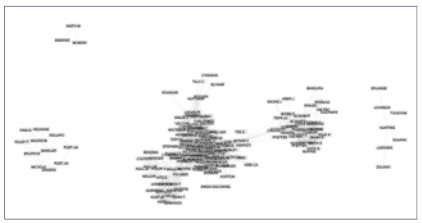
System evaluation requires further scenario development against example datasets. Domain experts with analytics needs are encouraged to contact the authors.

Contact Don Pellegrino (don@drexel.edu);
Chaomei Chen (Chaomei.Chen@ischool.drexel.edu).

Methods:

1. Identify substructures in the data. → 2. Report the substructures as candidate hypotheses and visualize them in context to support sense-making tasks. → 3. Evaluate the significance of proposed hypotheses through two dimensions
   a. Evaluate structural and topological properties of the hypotheses.
   b. Evaluate the content of the hypotheses.

1a.

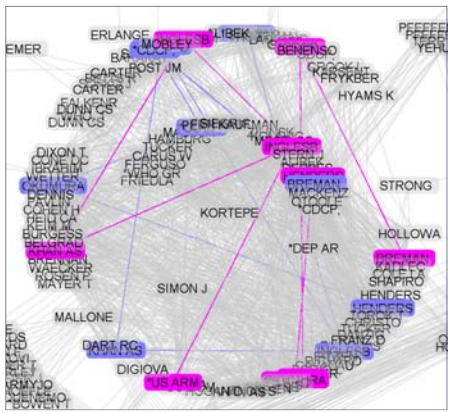


1b.

Example

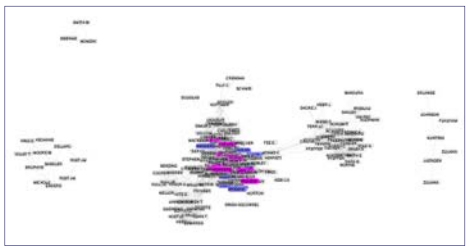
Visualization alone is not sufficient for identification of all potentially significant relations in a highly connected associative network. The left figures (1a and 1b) show author co-citation data from journal articles on terrorism. The network contains 198 nodes and 1909 edges displayed using a force directed layout (1a) and a radial tree layout (1b).

2a.



SemanticNetSA uncovers complex connections found within social network data; data for co-citations in scientific articles illustrate the potential. Two discovered substructure instances are highlighted with color in figures 2a and 2b. In this case both instances are found within the most densely populated areas of the visualizations. The structures in this example are relations between the co-citations of articles written by specific authors.

2b.



The identified authors are related in a complex way. Further, the complex relationship is acted out between them more than once. The repetitive complexity is sufficiently rare for this data that it is hypothesized it may indicate a sophisticated emergent phenomenon. An analyst may use contextual information from the high level view to decide if the relationship is worthy of investigation. If the central cluster is not of interest the relationship may be safely ignored. If the central cluster is being monitored then this data may warrant further investigation.

2c.

The analyst repositions the two instances to explore the hidden relationship (2c). Node selection reveals the evidentiary data used to formulate the hypothesized relationship. In this example papers from INGLESB, INGLESB and KHAN AS were respectively cited. One of the papers by INGLESB that was cited in that set was cited with US ARM and HENDERS. OKUMURA was cited with the HENDERS paper from that set. A CDCP paper was cited with the OKUMURA from that set. The CDCP cited with the one from OKUMURA was cited by another one also from CDCP. BREMAN was cited with the second CDCP article. This entire set of relations was repeated again for the same authors but completely different papers.