EuroVis 2016
Machine Learning Methods
in Visualization for Big Data
6 June 2016, Groningen, the Netherlands

Graph Mining & Integration of ML and Vis

Daniel Archambault[1]

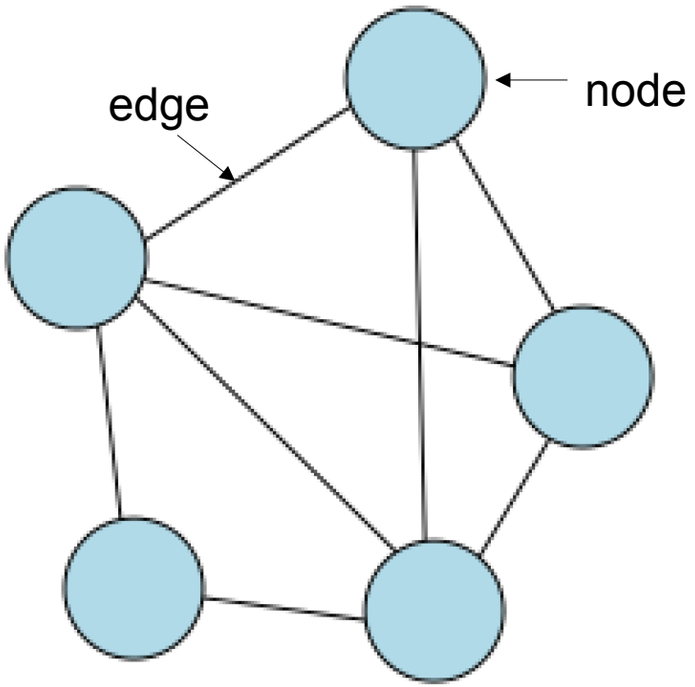
[1]Swansea University

Outline

- Introduction
- Community Finding Approaches
- Evaluation of Approaches
  - Normalized Mutual Information (NMI)
- Multivariate Graphs
- Discussion

What is a graph/network?

- Encoding of entities and their relationships
  - Entities are **nodes**
  - Relationships are **edges**
- Can be directed or undirected



Applications

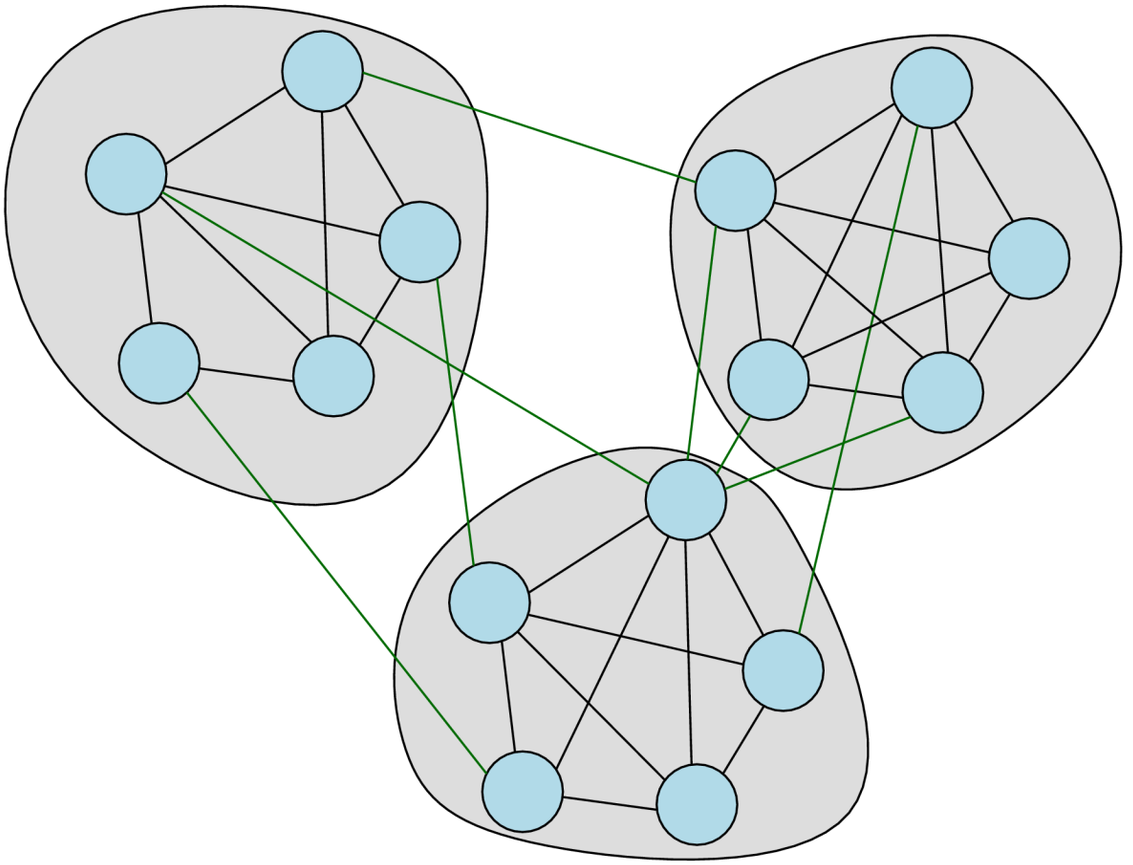
- Graphs have many applications
  - Social Networks (e.g. Facebook, Twitter, etc.)
  - Biological Networks (e.g. Gene/Protein interact)
  - Citation Networks
  - Computer/Software Networks
- Encoding provides a way to reason about higher order relations in this data

What is Graph Mining?

- Finding structure automatically in graphs
- Application of Data Mining to Networks
- Types of Graph Mining
  - Community Finding
  - Link Prediction
  - Subgraph Matching
  - ...
- Focus on community finding in this talk
- Relationship to clustering

What is Community Finding?

- Separate out graph into highly connected components
- Break few edges
- Cluster has strong connectivity



Why Community Finding?

- Identifies components that are highly connected
- In applications, these often mean something
  - Social Networks – social communities
  - Protien Networks – similar function
  - Citation Networks – fields of a discipline
- Highly connected components usually have meaning in network analysis
- Makes sense to detect them!

Community Finding Approaches

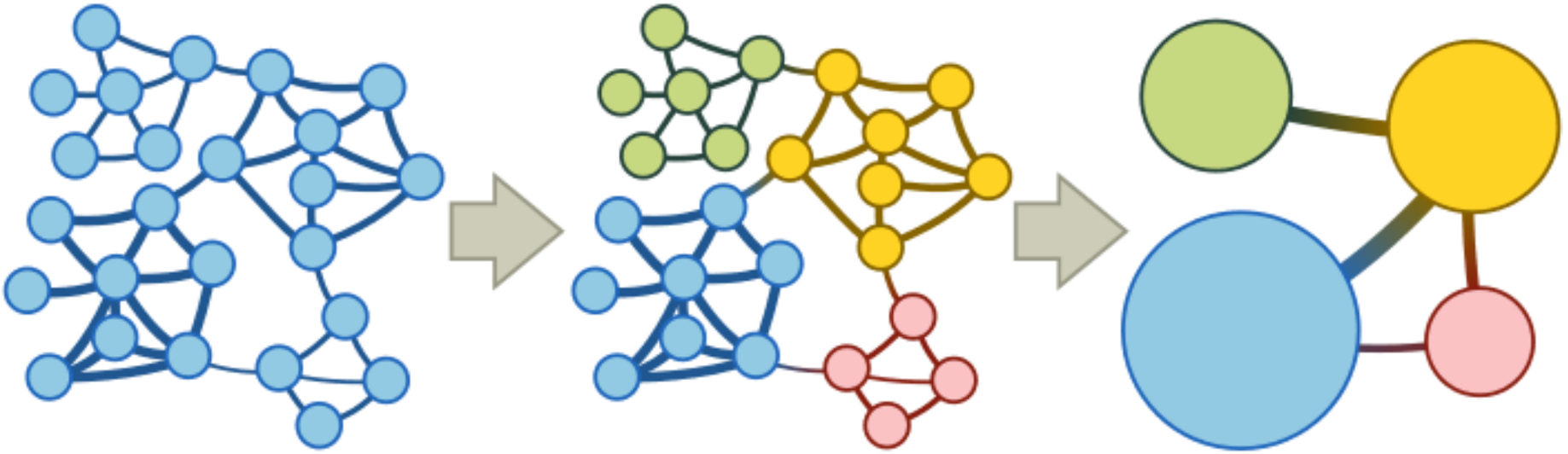
- Extensive research in community finding
- Many algorithms exist
  - commonly O(m) for m edges
- Examples:
  - Girvan & Newman 2004
  - Blondel et al. 2008
  - Palla et. al. 2005 (Cfinder)
  - **Rosvall & Bergstrom 2008 (Infomap)**

- Issue: Results not always deterministic
  - Get to this in a second...

How Does Infomap Work?

- Optimises division of graph into tightly connected components

http://www.mapequation.org



- It does this via probabilities, but there is a nice analogue via physical analogy

Random Walk Transmission

Video

http://www.mapequation.org

Community Finding Study

- Empirical study testing leading algorithms against each other

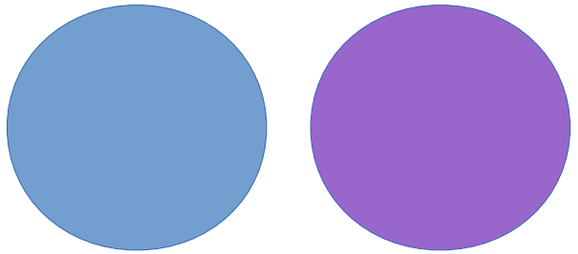
Andrea Lancichinetti and Santo Fortunato.
Community detection algorithms: A comparative analysis.
Phys. Rev. E 80, 056117, 2009.

- Experiment exhaustively testing community finding approaches by comparing them to known ground truth (LFR benchmark)

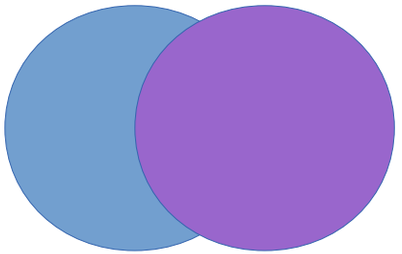
Evaluating the Output

- Normalized Mutual Information (NMI) is used to evaluate the similarity between two sets of communities.

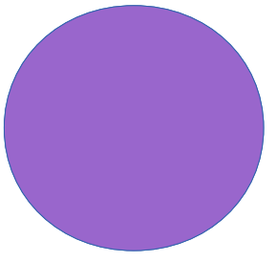
- Metric measure degree of match between the nodes in each community



No correspondance 0



Partial correspondance (0,1)



Perfect correspondance 1

Study Procedure

1. Generate community structure using LFR. This gives a graph and a correct answer.
2. For each algorithm, try and detect this community structure
3. Use NMI to compare the detected communities to the correct answer
   - The closer to 1 means the closer to the embedded ground truth

Study Results

- Infomap performed the best.
- Blondel et al. 2008 and Girvan & Newman 2004 also performing well

- In addition the study tested random graphs, where there should be no community structure, and found these algorithms performed well in this circumstance

Stability Issues

- Community finding approaches require random seeds
- Therefore, different outputs could occur for the same run of the program
- A solution: report the **average** community structure
- This is known as consensus clustering

- *Andrea Lancichinetti and Santo Fortunato. Consensus clustering in complex networks. Nature Scientific Reports 2 (336).*

Human Centred Results

- Similar results found from a human centred perspective

Alexandra Lee and Daniel Archambault. Communities Found by Users -- not Algorithms: Comparing Human and Algorithmically Generated Communities. ACM Conference on Human Factors in Information Systems (Note, ACM CHI 16), 2396-2400, 2016.

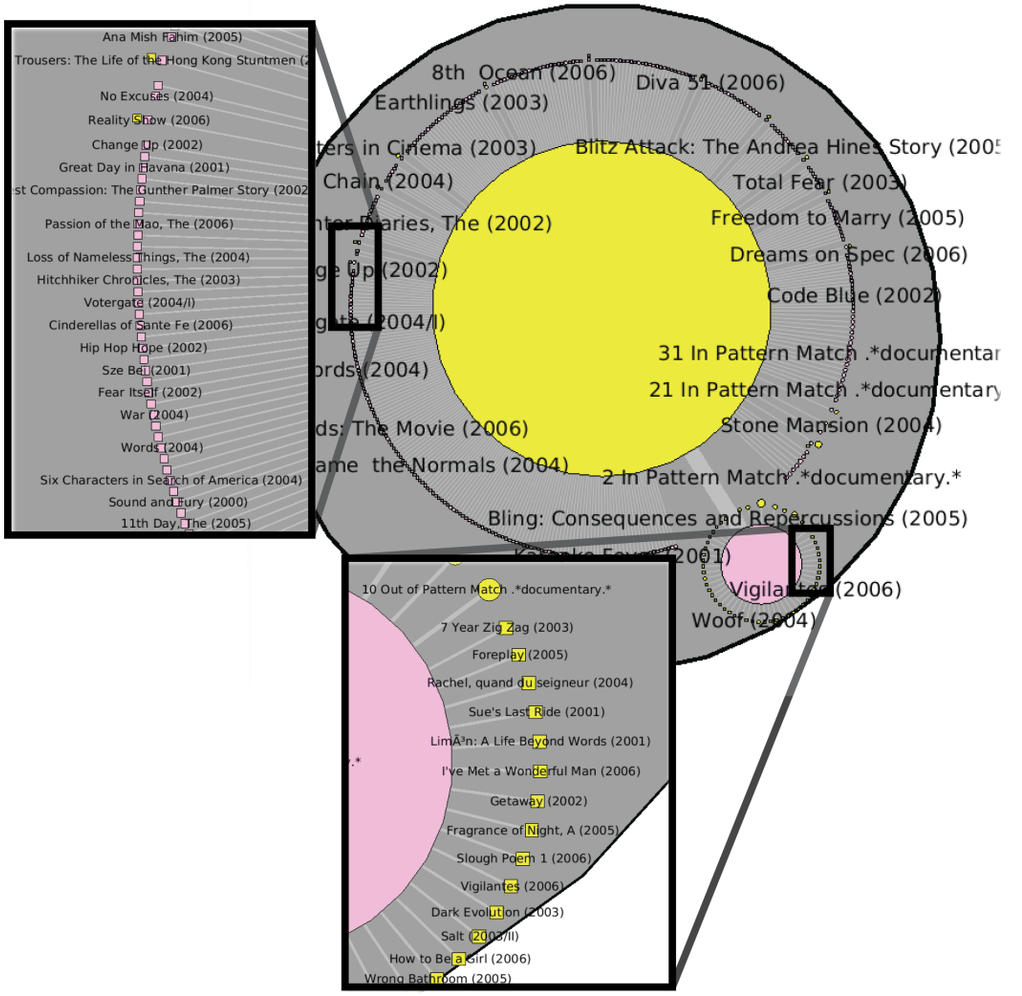
- Study compared human annotated communities with automatically found ones

Multivariate-Based Visualization

- Early work on visualization methods for multivariate graphs
  - ASK-Graph View and GrouseFlocks
  - TugGraph
  - Semantic Substrates
  - Pretorius thesis

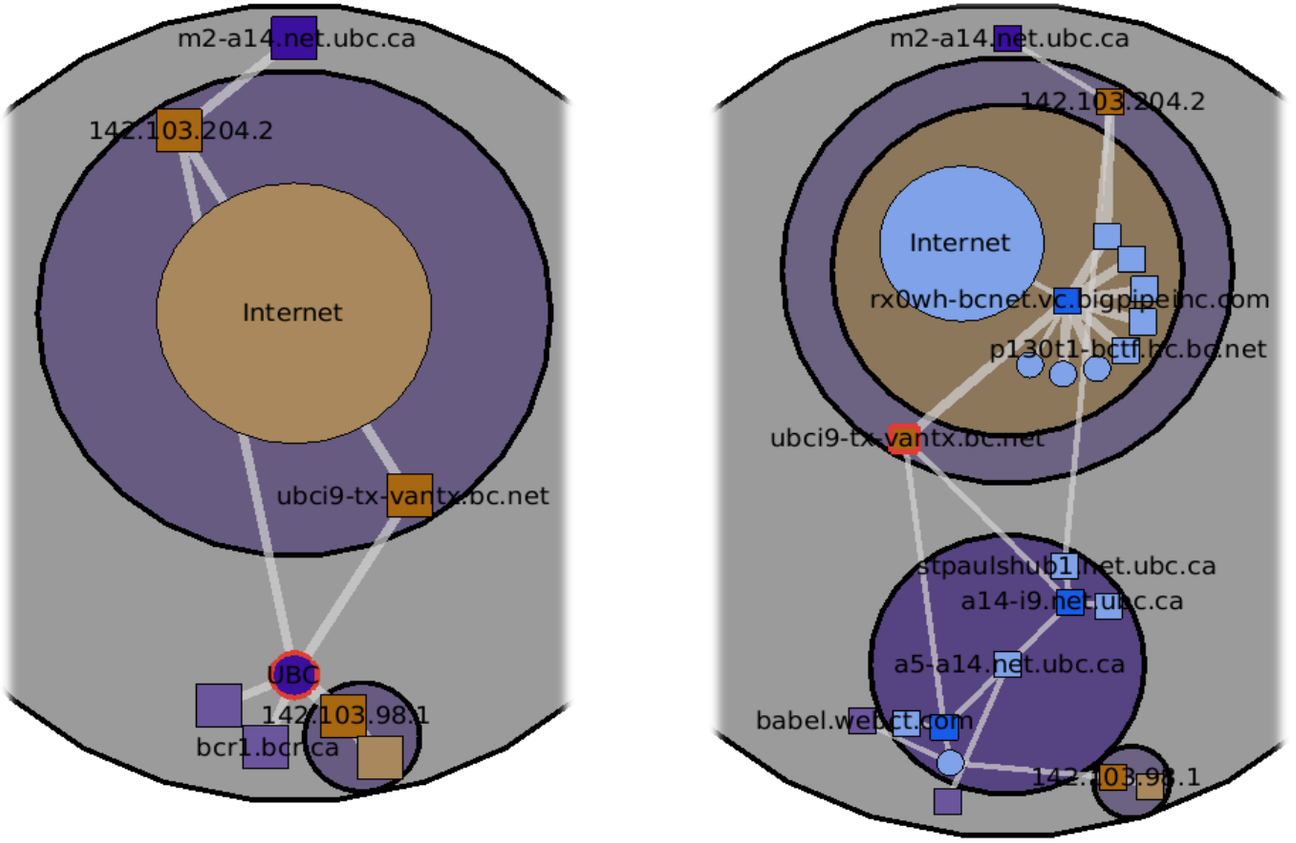
ASK-Graph and GrouseFlocks

- Visualization method for large clustered networks
- Attribute driven clustering and visualization of networks
  - Draw clusters on demand



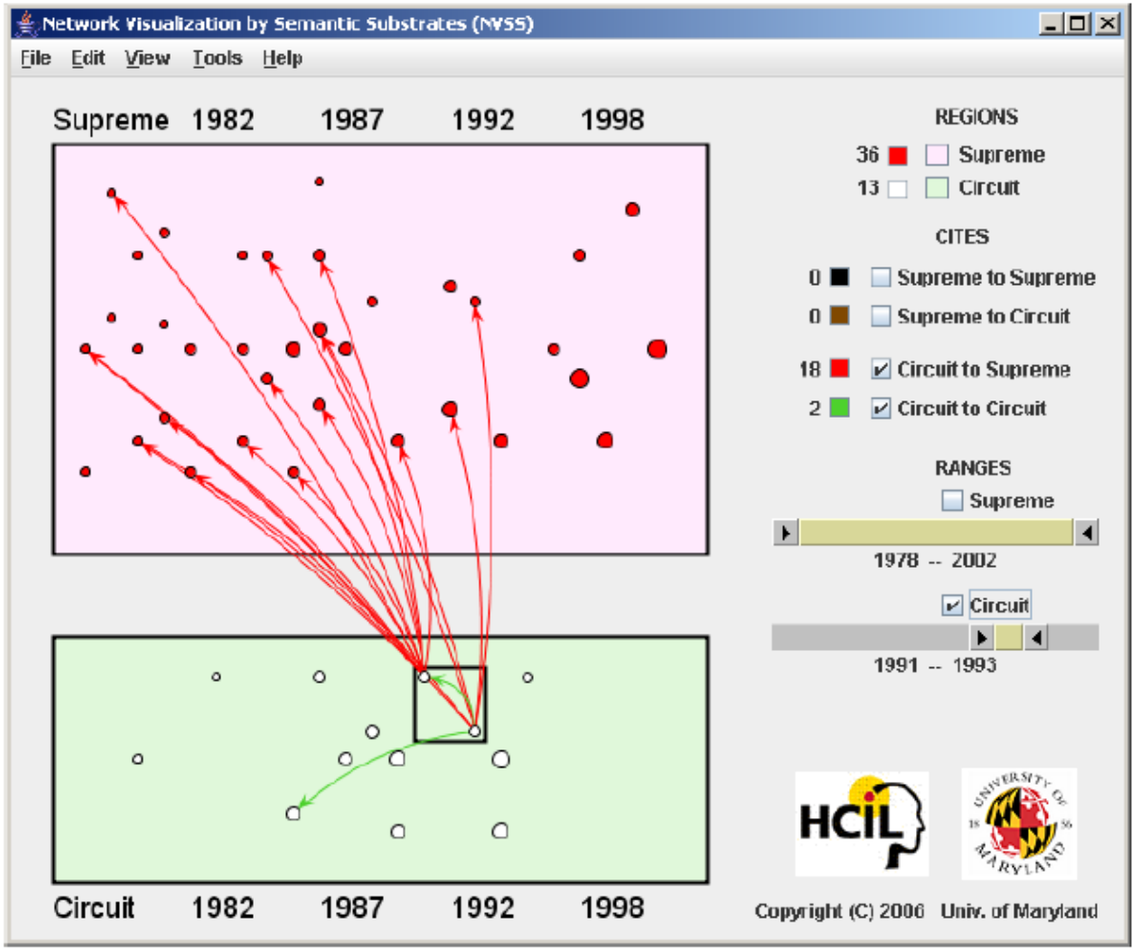
TugGraph

If interested in the area around a node or component can tug out structure nearby



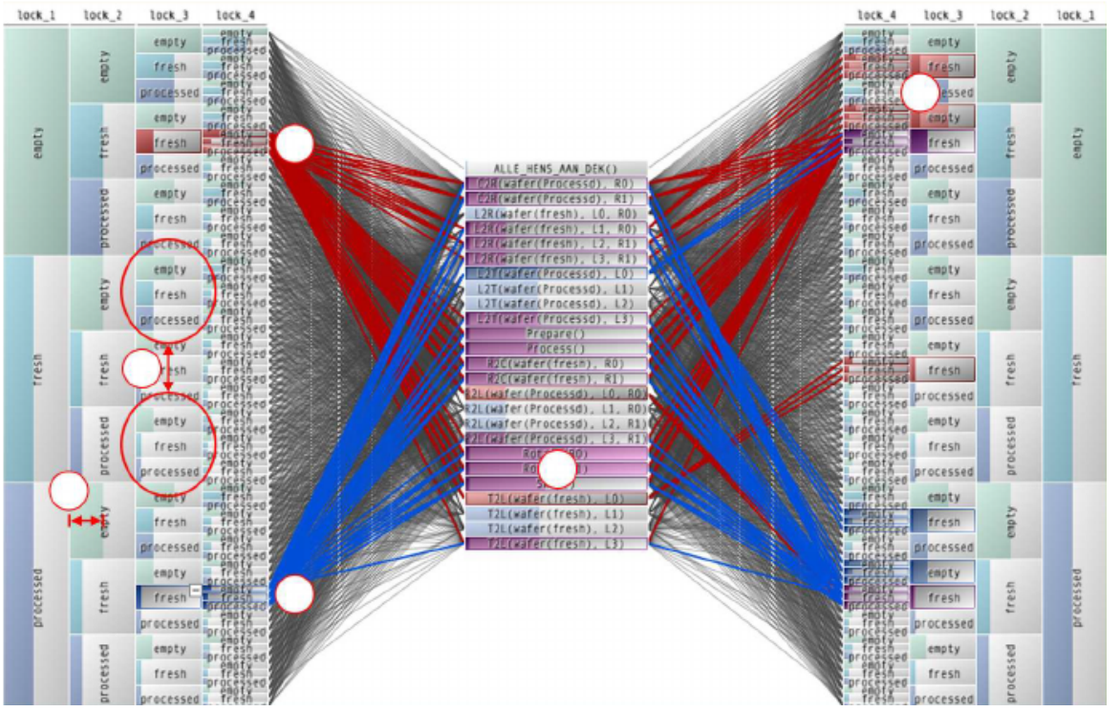
Semantic Substrates

Network visualization where spatial position encodes attribute values



Pretorius *et al.*

- Extensive work on multivariate and state transition graphs
- EuroVis 2008 paper on multivariate graphs is especially interesting

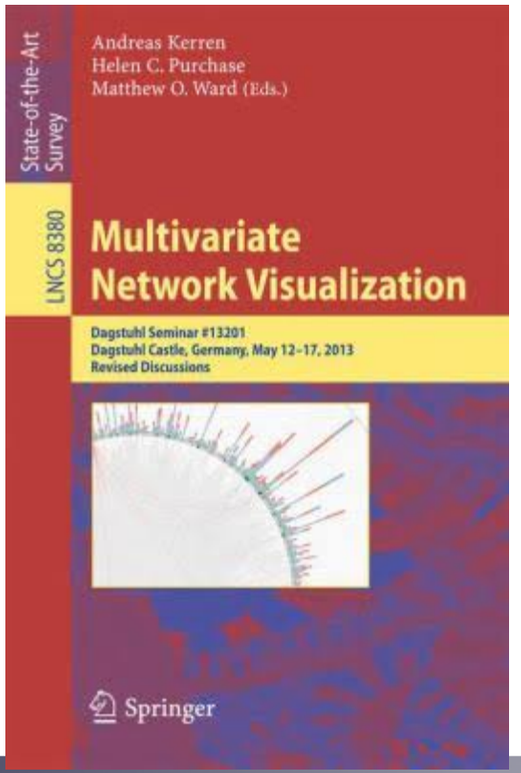


Book on Multivariate Graphs

Springer book on this topic as the result of a recent Dagstuhl workshop

http://www.springer.com/us/book/9783319067926

.



Relevant Surveys

Very nice survey on graph visualization:

von Landesberger, T., Kuijper, A., Schreck, T., Kohlhammer, J., van Wijk, J.J., Fekete, J.-D. and Fellner, D.W. (2011), Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges. Computer Graphics Forum, 30: 1719–1749.

Recent STAR on Dynamic Graphs:

Fabian Beck, Michael Burch, Stephan Diehl, and Daniel Weiskopf. The State of the Art in Visualizing Dynamic Graphs. In Proceedings of State-of-the-Art Reports of EuroVis 2014.

NMI Software and Community Finding

Link to Infomap Community Finding Algorithm:
http://www.mapequation.org/code.html

Link to Normalized Mutual Information Code:
https://github.com/aaronmcdaid/Overlapping-NMI

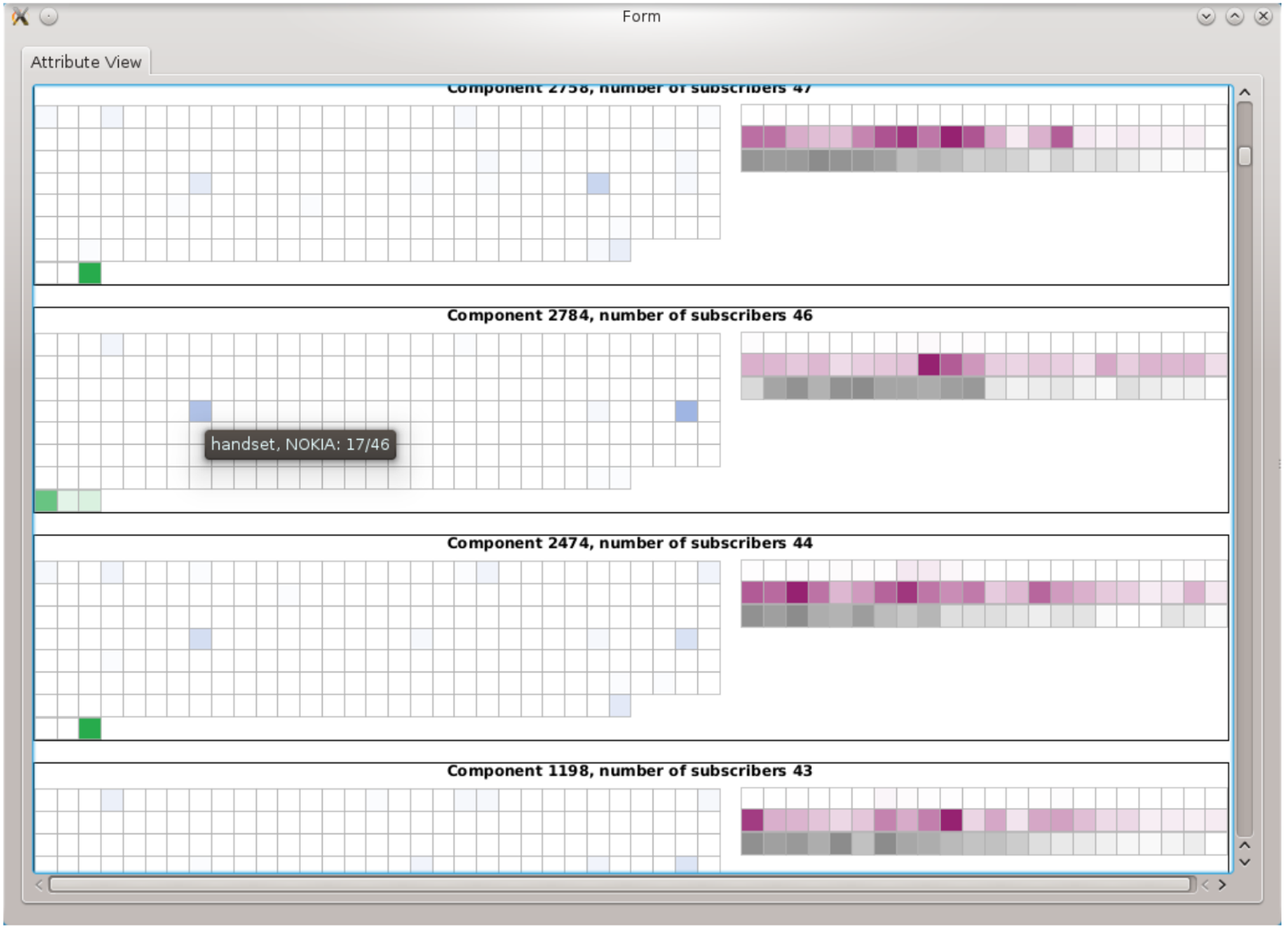
Full Fortunato Survey:
https://arxiv.org/abs/0906.0612

Visualization Meets ML

- Over the course of the day, we have explored many different techniques for automatically finding patterns in data
- In this room, many of us are visualization experts
- We are only beginning to determine ways which visualization and machine learning can work together.
- Mostly going to concentrate on my experience

Example Churn Analytics

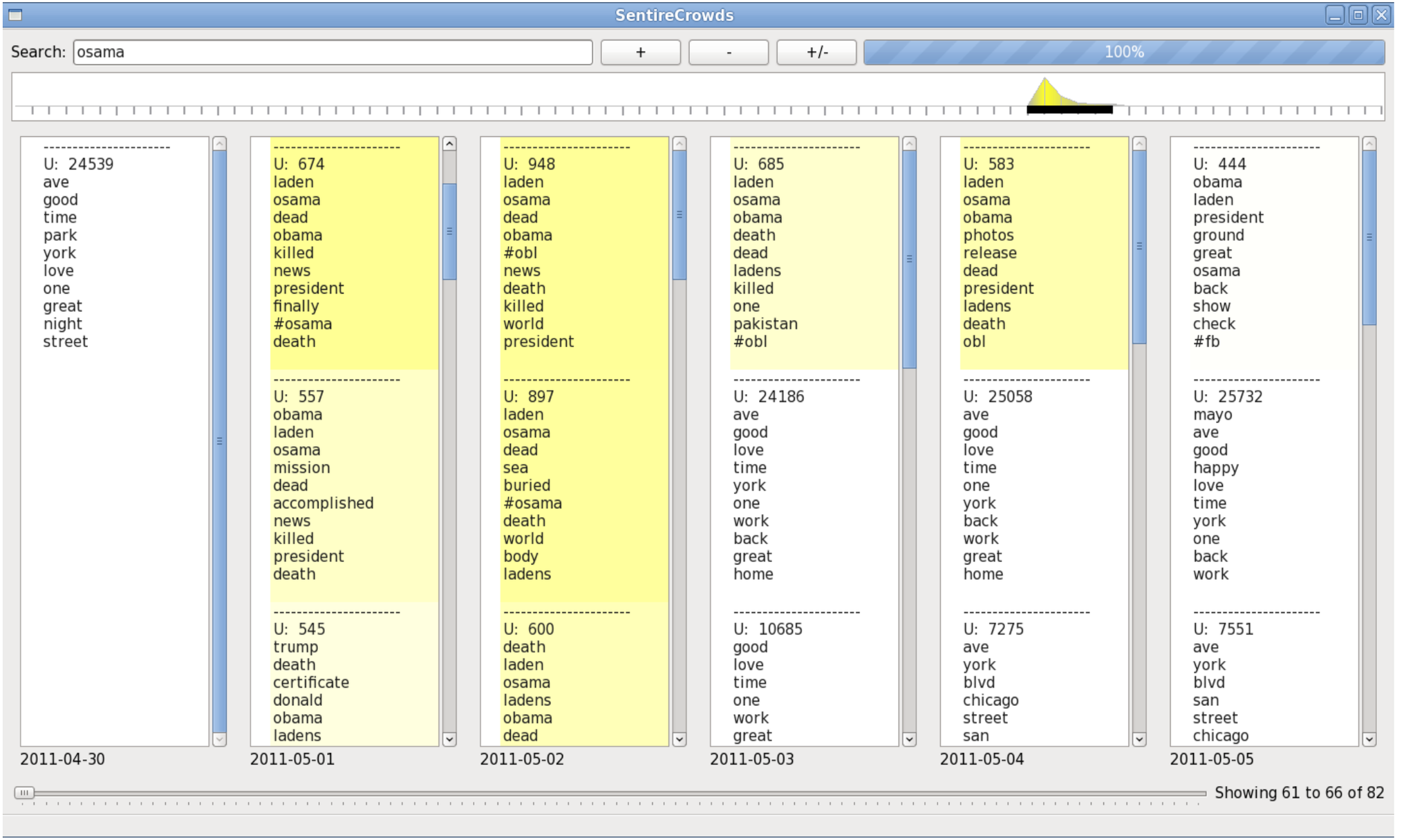
- Very large graph of nearly 1 billion edges
- Summaries of components enriched in churn



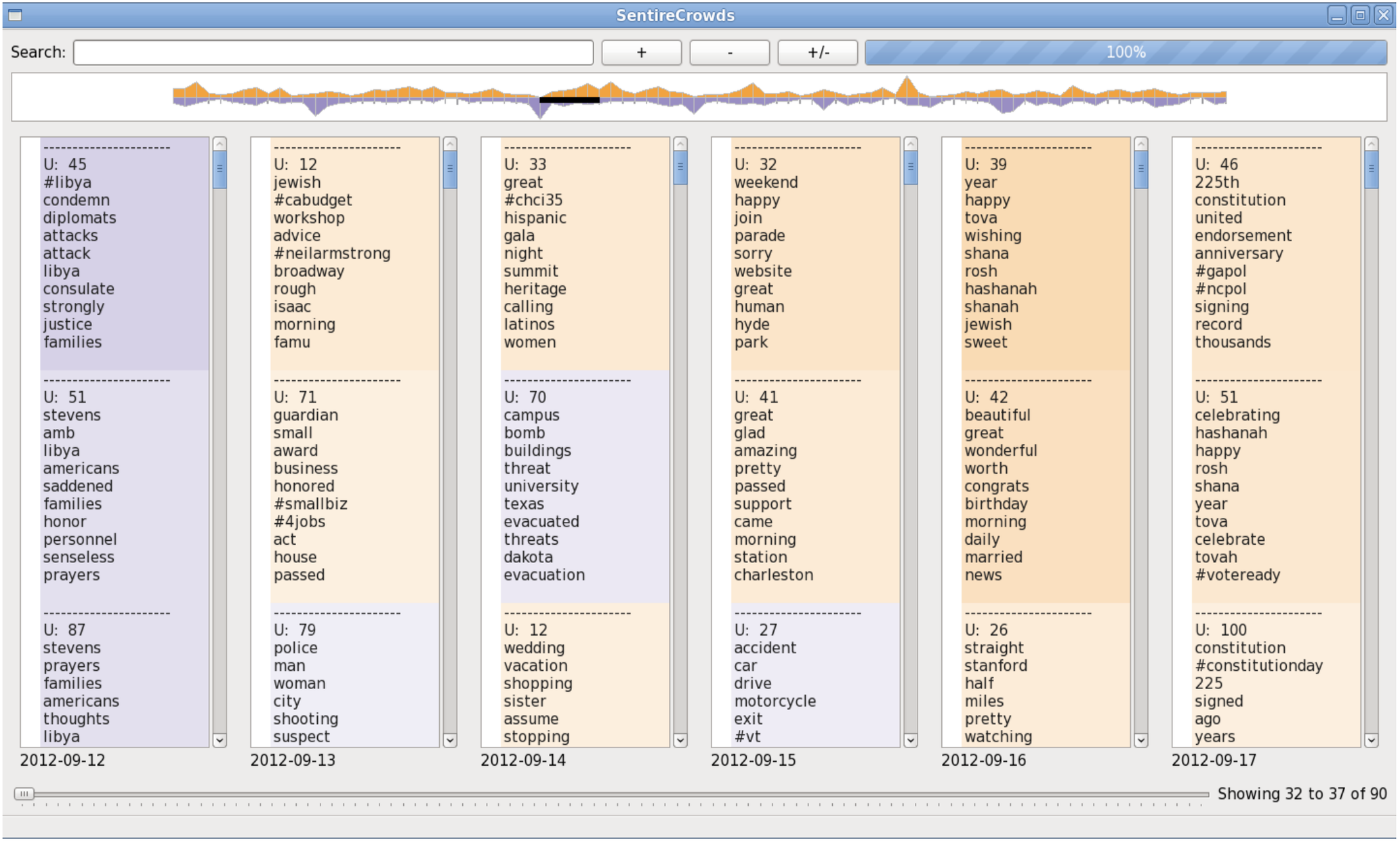
Twitter Analysis

- How do you look at tens of millions of Tweets?
- Worked with members of a network analytics and data mining group to create a dashboard for navigating these tweets.
- Discover areas enriched in a topic or highly positive and/or negative.

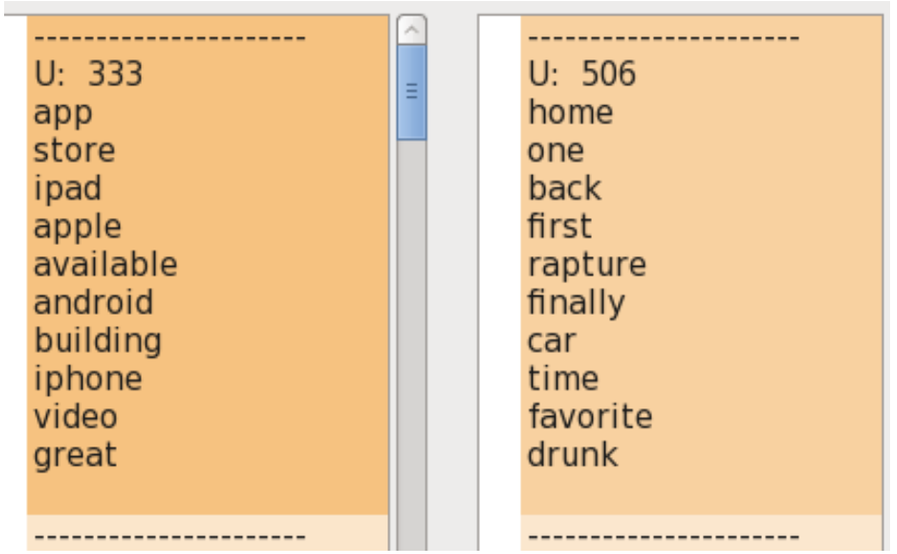
Example Twitter Analysis



Example Twitter Analysis



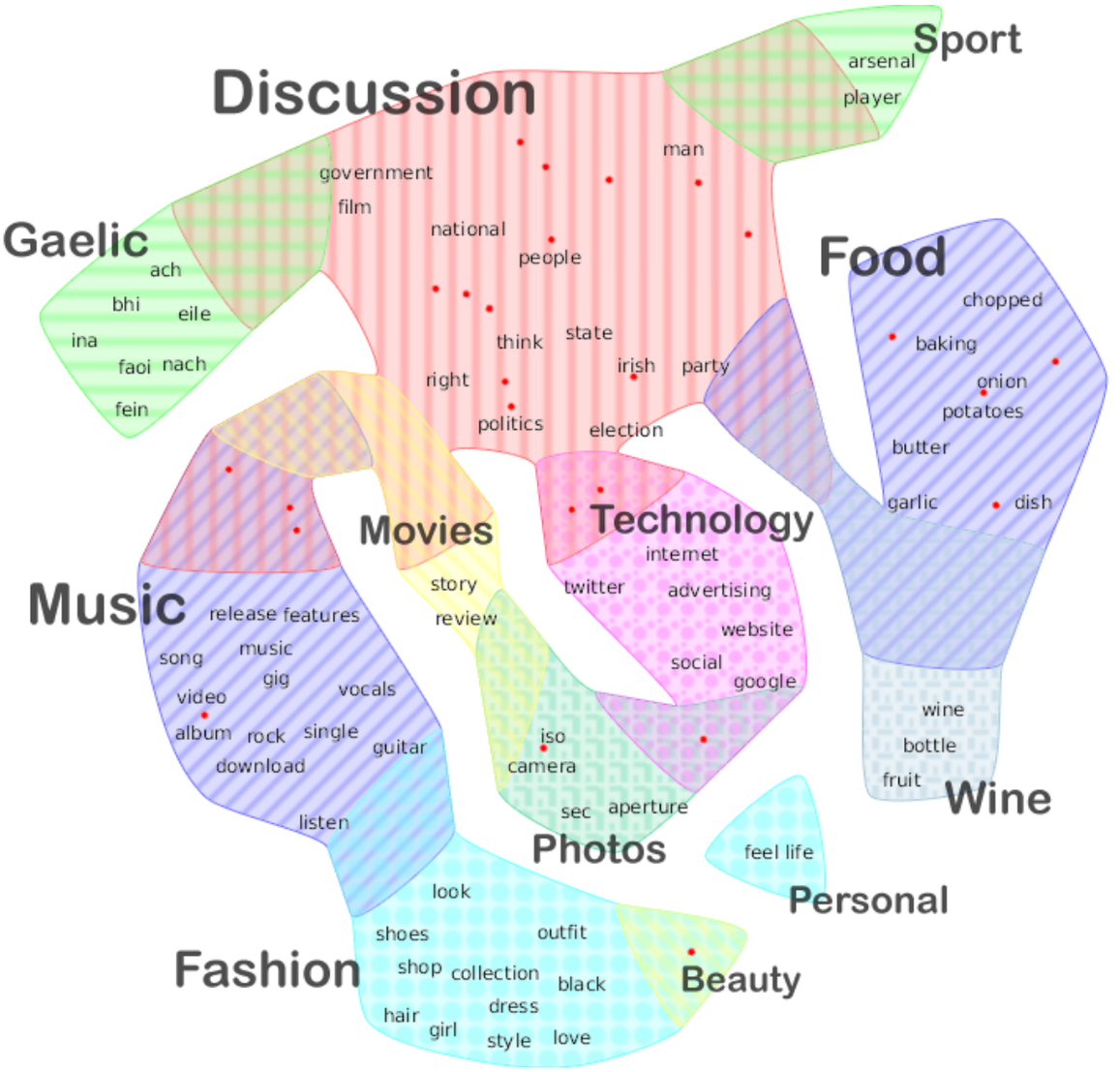
Expect the Unexpected



- What is going on here?
- Why are people positive about these topics?

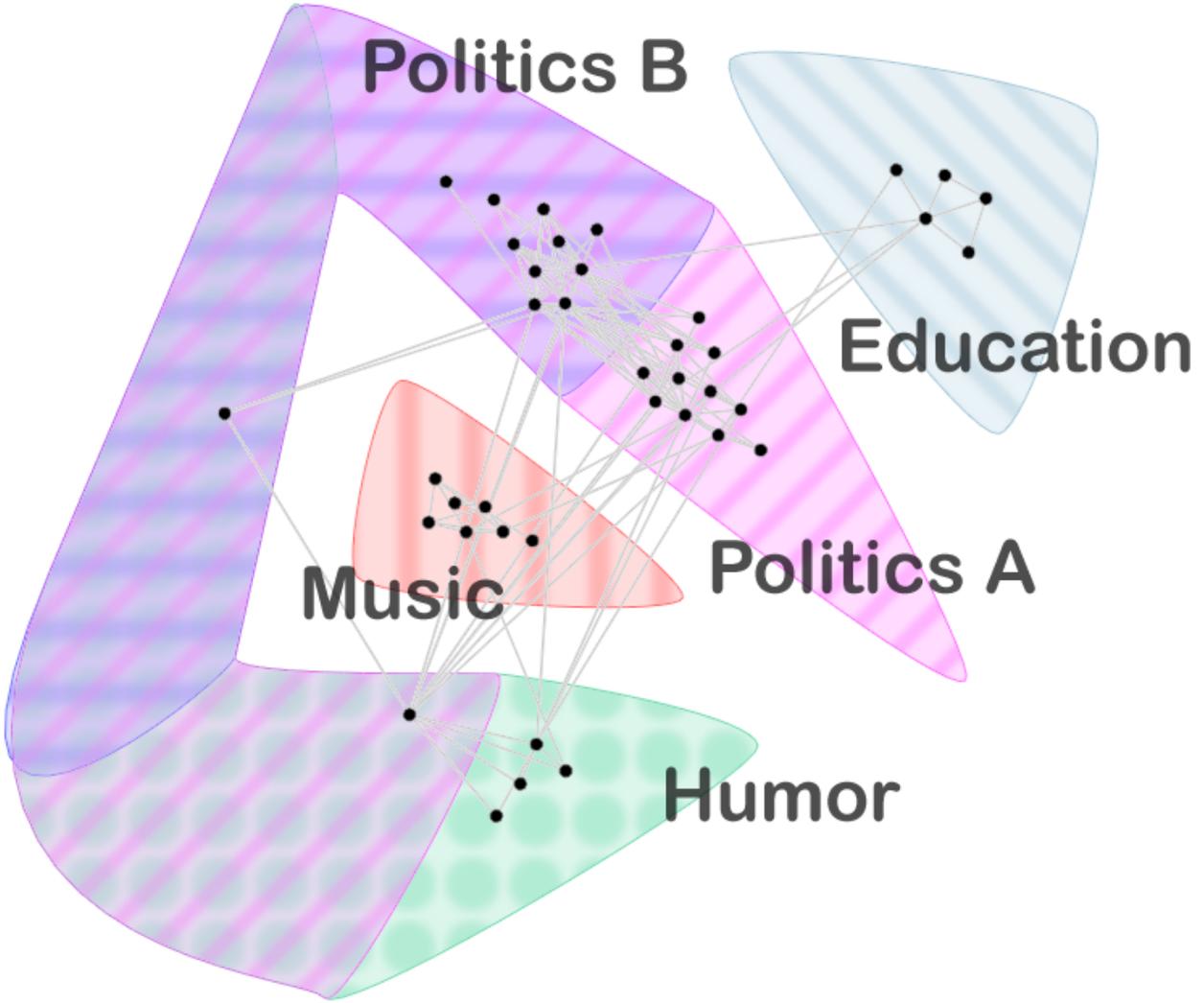
Blog Analysis

- PhD Student in English asked what does the Irish Blogosphere look like?
- Text perspective of language used



Blog Analysis

- Decomposition of discussion via link structure



Blog Analysis

- Recommendations to English researcher

Theme	*Representative Blog*
Beauty	** beaut.ie
Education/Law	** cearta.ie
Fashion	blanaid.com
Food	** icanhascook.wordpress.com
Gaelic	miseaine.blogspot.com
Humor	counago-and-spaves.blogspot.com
Movies	scannain.com
Music	** irishtimes.com/blogs/ontherecord
Personal	anonomousangel.wordpress.com
Photos	slkav.com
Politics	splinteredsunrise.wordpress.com
Sport	dangerhere.com
Technology	** mulley.net
Wine	firstpress.blogspot.com

Discussion

Reflecting on today's activities, how can our two fields better collaborate? What avenues of research do you feel are the most fruitful?