Graph Mining and Visualisation

Daniel Archambault

Previously in CSCM27...

- What is stability? When does it help?
- Name some integrated visualisations for dynamic graphs

Previously in CSCM27...

- We now look at the interplay between graph mining and visualisation
- And it's the last lecture of M27

Mining and Visualisation

Dynamic Representations of Graphs

- Graph mining automatically finds features in the data
- Graph visualisation involves the users to make discoveries
- Interplay between the two of them makes a visual analytics system

Examples of Integration

- Look at two systems where there is interplay
  - first is based on large static graphs
  - second is a collaboration with English on language usage

A problem...

- In data science, the data sets we can collect grow larger
  - no different when it comes to network data sets
- Given a large graph, we turn it into a small graph...?
- What ways can we automatically turn it into a small graph?

Filtering is one way...

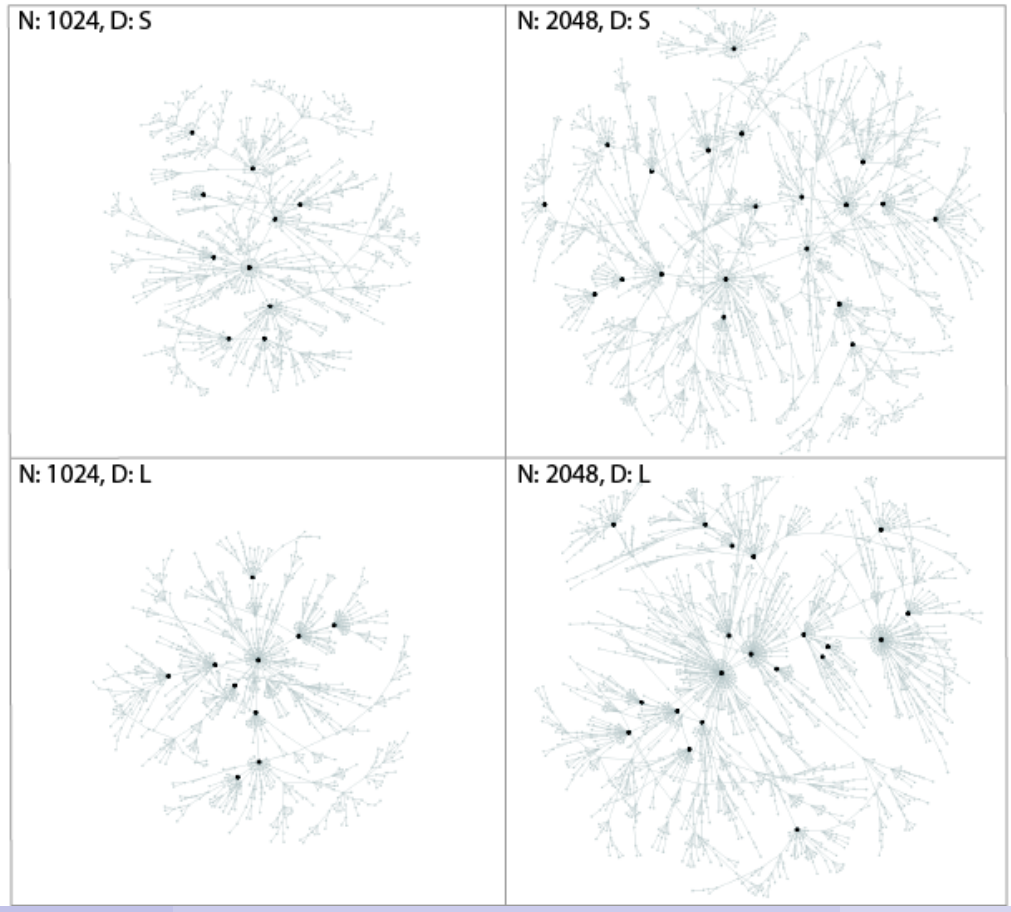
- Automatic filtering is sampling
- Many sampling methods exist in graph mining
  - Randomly select nodes
  - Randomly select edges (insert edges if both nodes in)
  - Select on a random walk
  - Select on a random walk with jumping
  - Forest-fire sampling (“burn” edges with probability)
- People are going to use these methods
- What are the effects on visualisations?

Experiment to test the effect

- Part 1: pilot to determine which factors important
  - high degree nodes, cluster quality, coverage
- Part 2: run three experiments to determine best performance
- Human observers do tasks on the data which are measured

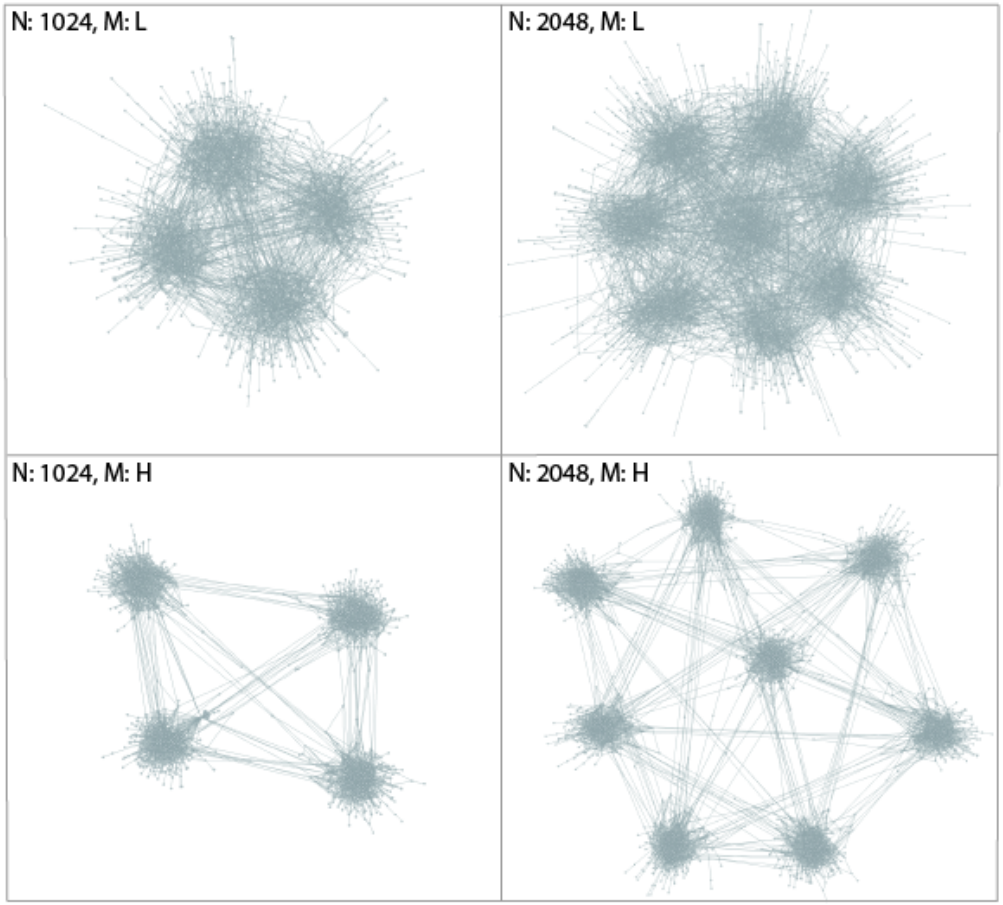
Is a high degree node still high degree?

- Given a sampled, do the high degree nodes still appear high degree?
- Random walk performs the best in this case



How are clusters preserved?

- Which method best preserves clusters
- Random walk with jumping and random edge



Is the entire graph covered?

- Are bits of the graph missing from the visualisation?
- Random walk with jump and random edge perform best
- Random walk has issues getting stuck
- Random node doesn't have great performance

Summary

- The data analytics method used influences the visualisation
- Not only in the “different data shown” sense
- But, also in the sense of perception
- It is important to take these things into consideration

An Examination of the Irish Blogosphere

- Social network perspective of the Irish blogosphere
  - *Identifying Representative Textual Sources in Blog Networks* (AAAI ICWSM 2011)
- *'Our local sphere' examined texts from the blogs that were identified representative and/or most influential within these Irish blog networks*

Motivation from a Humanities Perspective

- Collaboration with Karen Wade (Dept. of English)
  - English language usage in Ireland online
- From a humanities perspective, our methodology helps solve two principal problems:
- Practicality
  - The number of blogs that can be studied in detail by any one researcher is limited.
- Precision of selection methodology
  - Some previous studies into blogging within the humanities and social sciences have produced conflicting/confusing results.
  - With random or researcher-biased samples, there is a risk of reinscribing problematic assumptions about blogging.

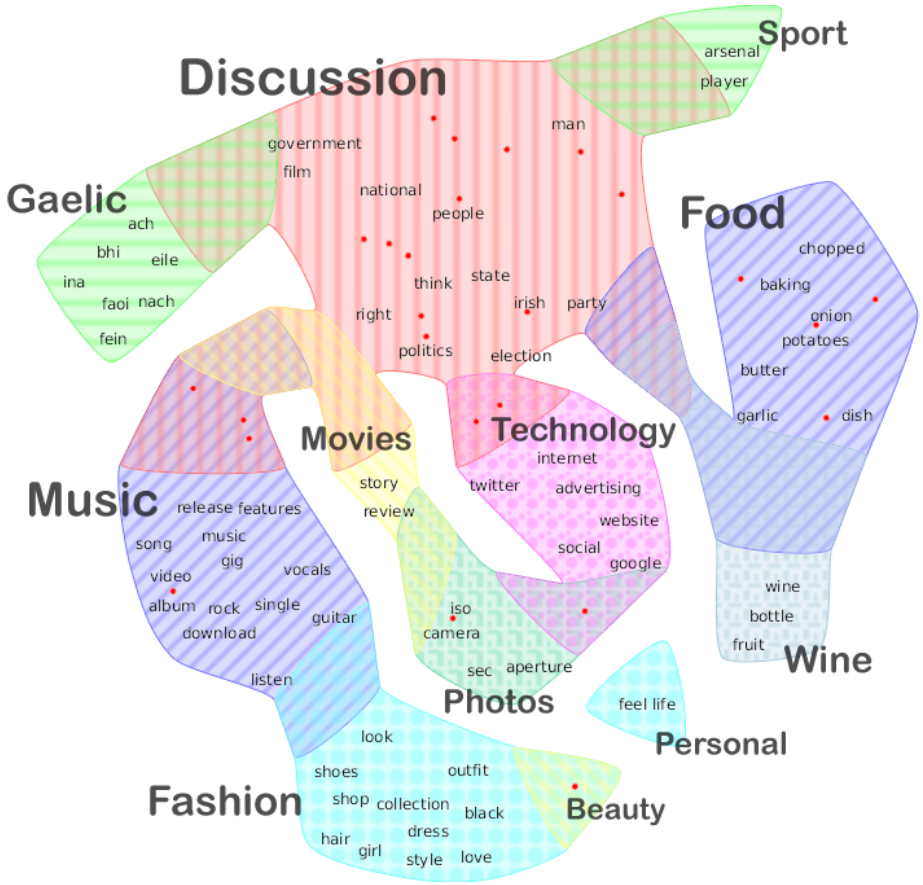
Data Set

- The Irish Blog Awards (2010)
  - Provided an initial sample of peer-recognized Irish bloggers
- From this seed set...
  - identified a set of 635 Irish blogs.
  - collected all posts for the period 1997-2011 (179k unique posts)
  - collected blogroll and post-link networks

Finding Central Blogs

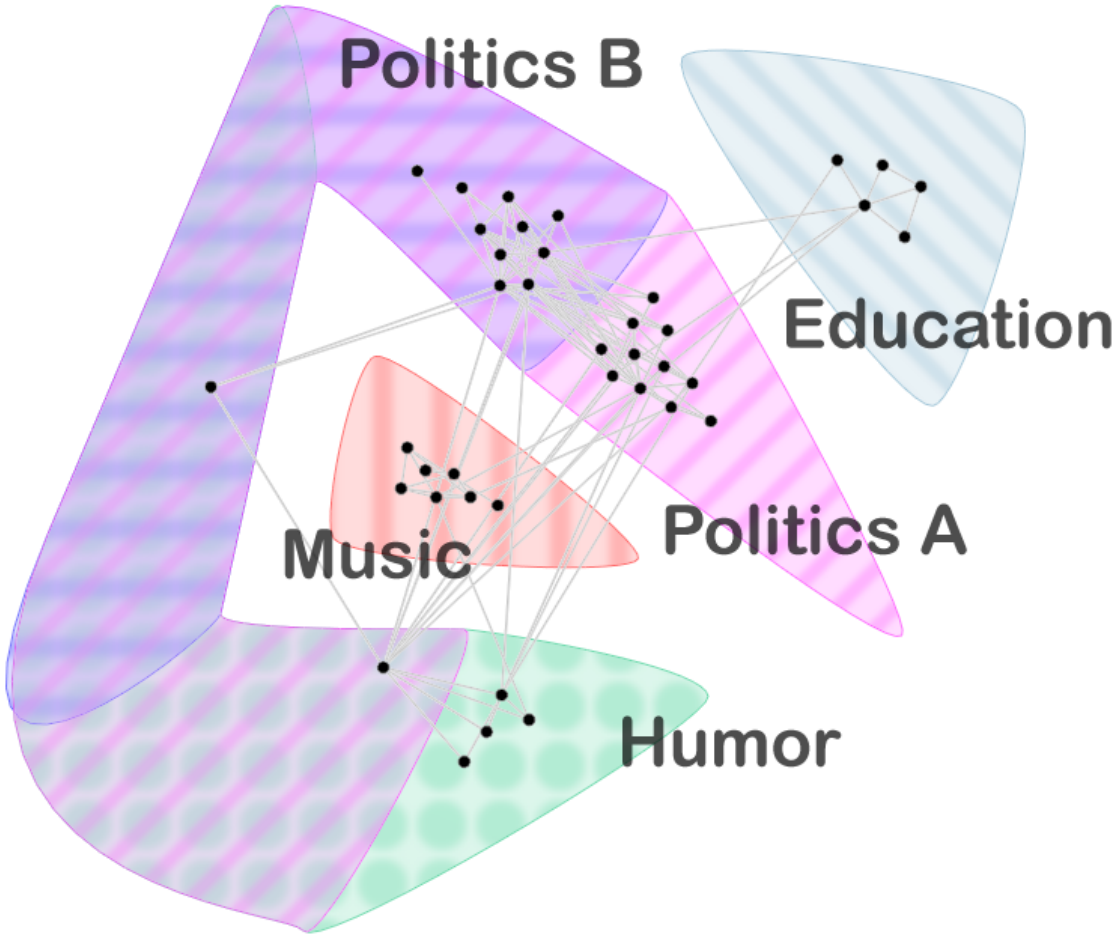
- Initially, in-degree network centrality used
  - identified well known members of the blogging community
  - did not provide good coverage
  - topics of these blogs were mostly technology and web
- How can we get a more representative sample?
- Overlapping Cluster Analysis
  - concatenate the all blog post text for each blog (profile)
  - calculate frequencies of all terms in the concatenated text
  - cluster profiles (NMF clustering) to identify topical groups

Text Similarity Visualization



- Similarity of text yields overlapping structure
- Topics are identifiable from keywords
- Covers 419 (68%) from the full set of blogs

Decomposition of Discussion via Links



- Discussion decomposed by link structure
- Small community (NI) and larger Irish community

Recommended Blogs and Conclusion

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

- Recommended blogs based on analysis
- Blogs indicated by ** are high degree nodes
- Process conducted in a way that is valid in the humanities

FIN

- This is the end of the course material for the module
- But just the beginning for some of you
- I hope you enjoyed
- Best with the project and have a good break afterwards