Graphical representations of a corpus, and clustering on graphs

University of Birmingham, 11 February 2016

Yves van Gennip (University of Nottingham)

Joint work with R. Carrington, A. Hennessey, M. Mahlberg, S. Preston, K. Severn, V. Wiegand

Outline

- Co-occurrences of words
- Graphs
- Clustering
- Example: the Dickens corpus
- Conclusions and open questions

Word co-occurrences

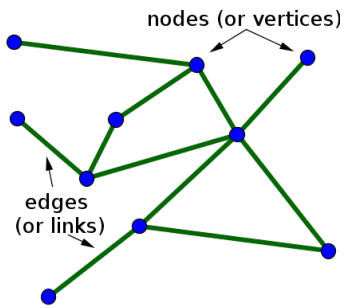
Word A co-occurs with word B if it appears near word A in the corpus. Choices:

- Window size: What counts as 'near'?
- Directionality: Do we make a difference between B appearing to the left or to the right of A?
- Weighting: Do we take into account how often B appears near to A? Does B appearing close to A count as a 'stronger' co-occurrence than B appearing further away from A (but still in the window)?

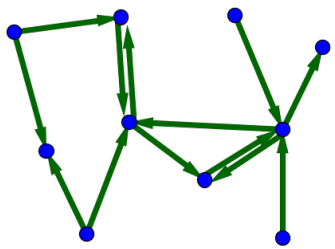
All these choices combined give us a pairwise relationship between all the words in the corpus.

Graphs

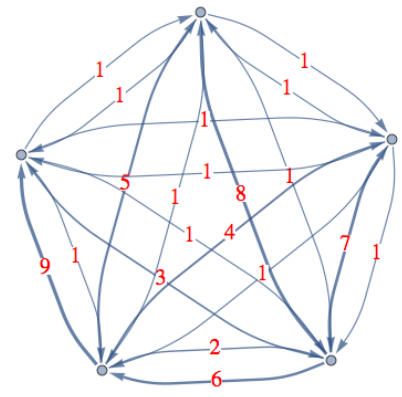
In mathematics the word 'graph' occurs with two different meanings: the graph of a function and a graph representing a network. Here we will be talking about the second meaning.



(a) Undirected graph



(b) Directed graph

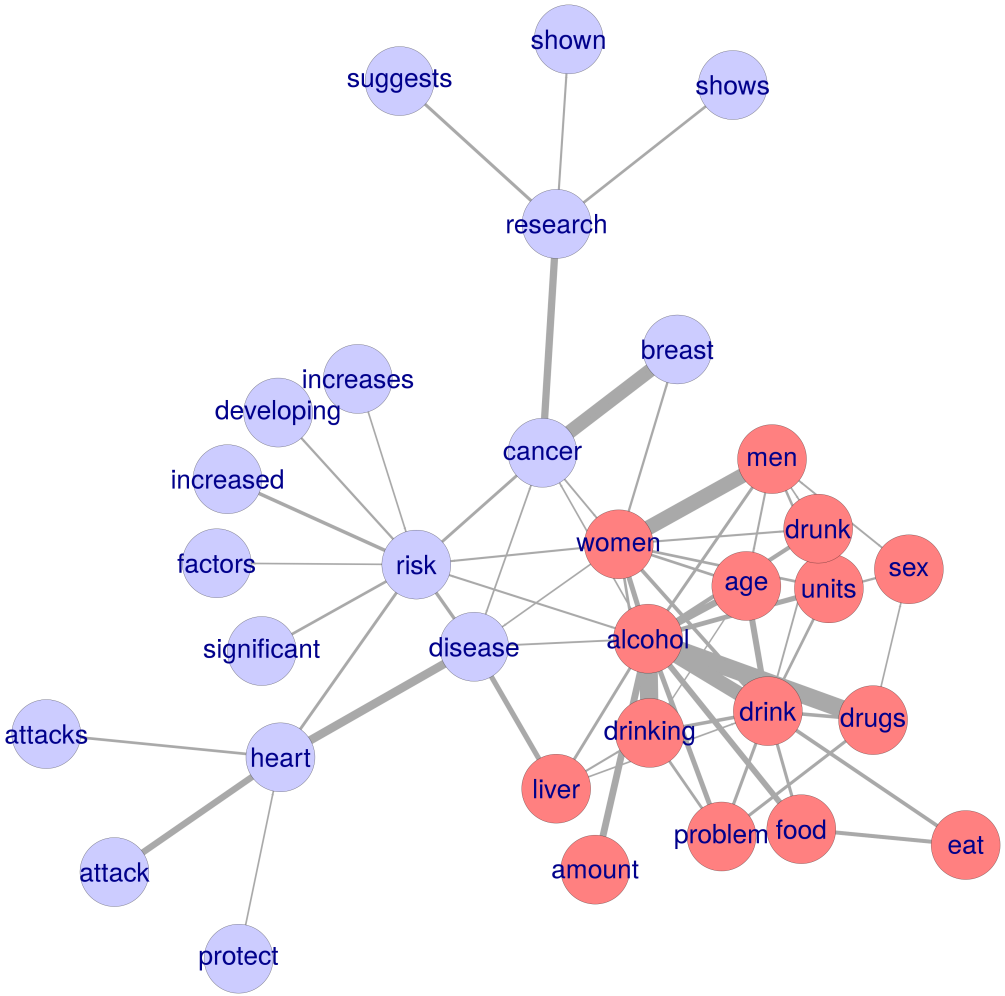


(c) Directed and edge weighted graph

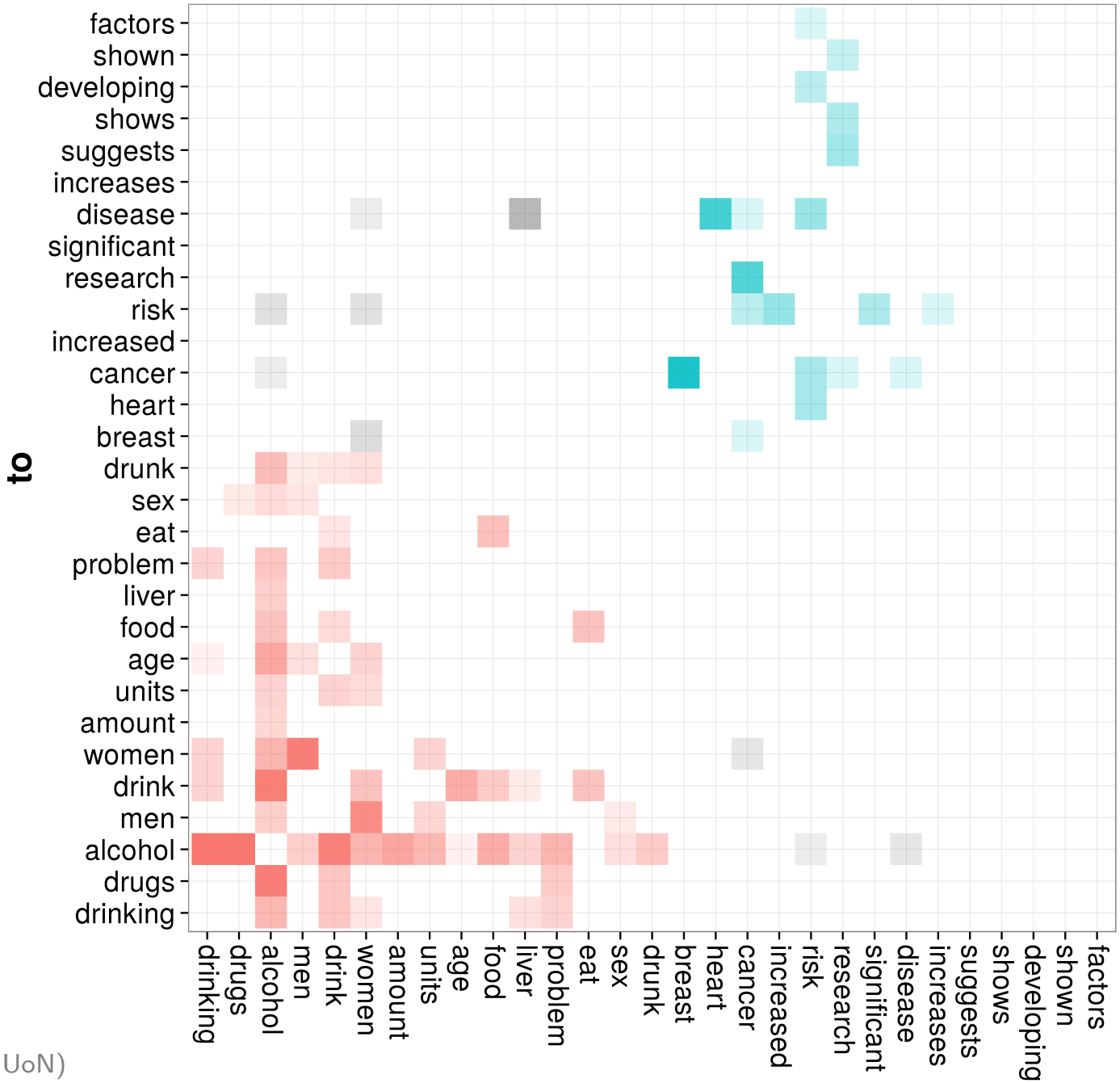
(a) & (b) Nykamp, DQ at Math Insight mathinsight.org

(c) DavidC at mathematica.stackexchange.com/questions/4520

Undirected co-occurrence graph

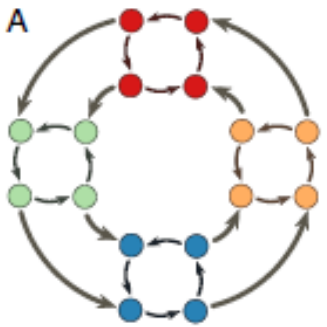


Undirected co-occurrence graph: matrix representation



Clustering

Graph clustering aims to group vertices together in a meaningful way.
What is 'meaningful' depends on the structure of the graph and the context in which clustering happens.
There are many different methods for graph clustering. The examples in this presentation were done using the *infomap* algorithm which works on weighted directed and undirected graphs.



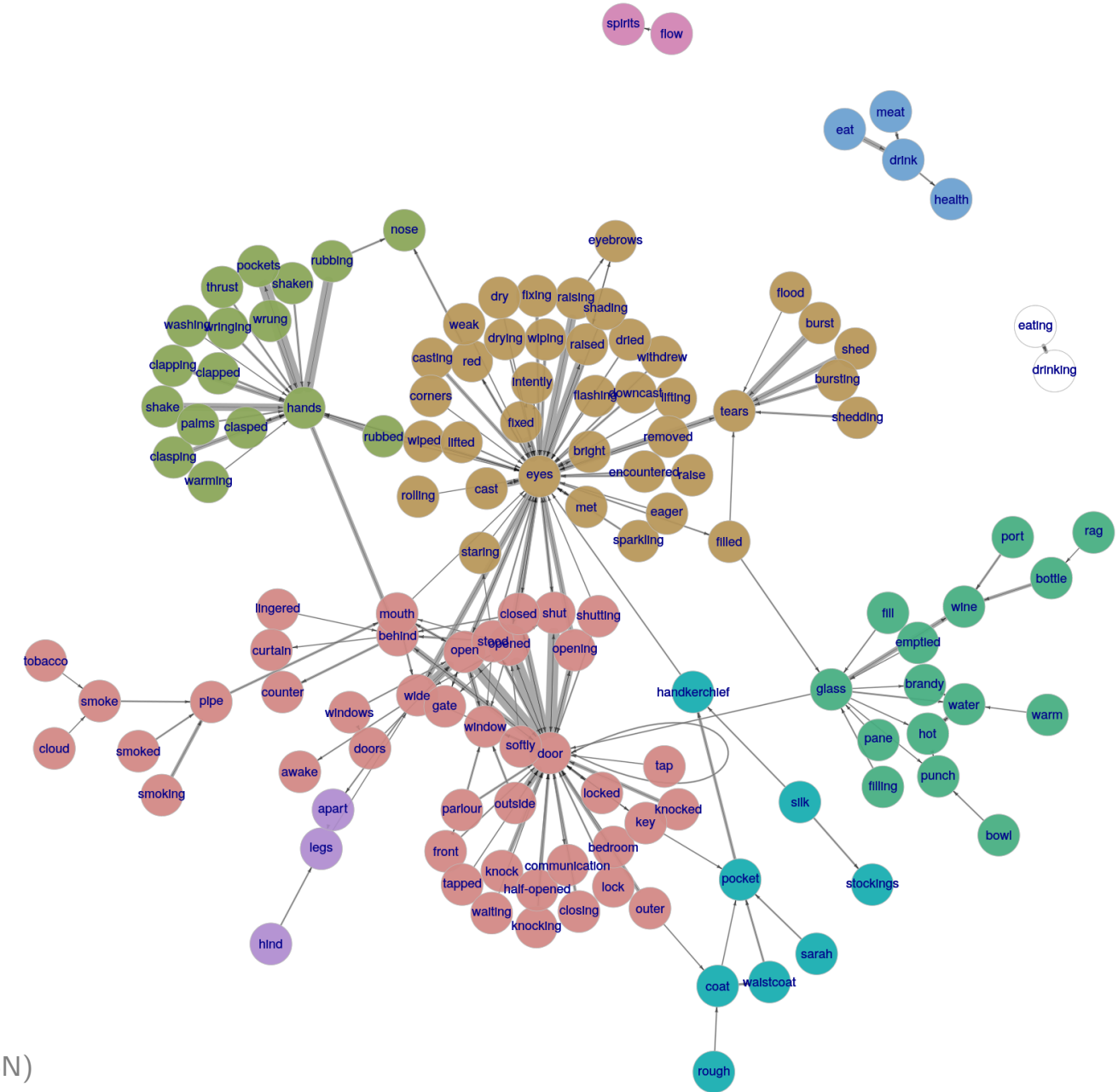
Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure.

Proceedings of the National Academy of Sciences, 105(4), 1118–1123.

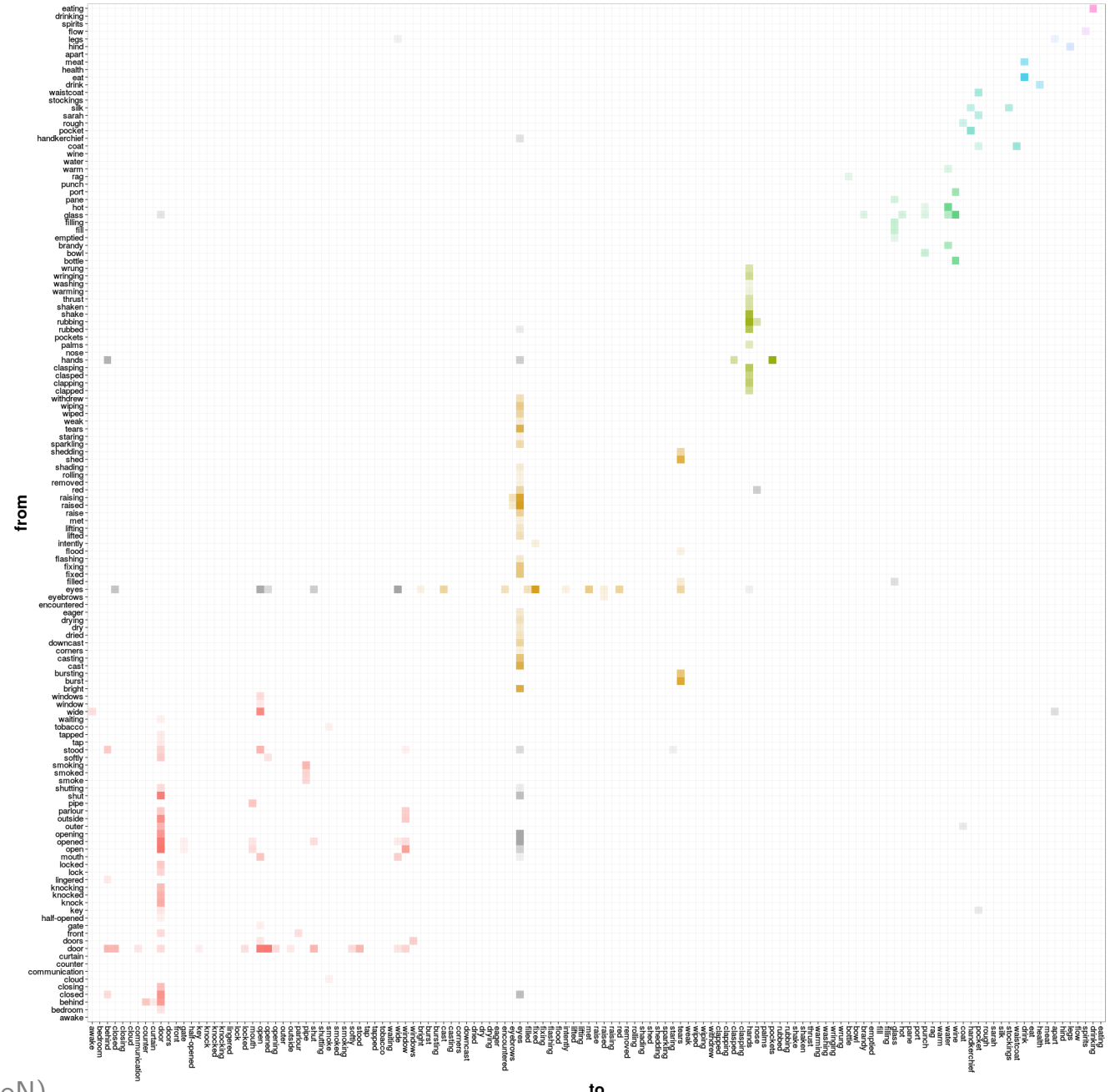
Example: the Dickens corpus

- All 15 novels in the Dickens corpus
- Tokenizing is done using the R native scan function
- Window: Using counts directional co-occurrences with a span of 4
- Filter for visual display: only include co-occurrences that appear more than 10 times
- Clustering using the infomap algorithm
- During clustering we ignore any co-occurrences which include any of a list of stopwords (otherwise the clusters form around the stopwords)
- For visualisation: focus on clusters containing a set of search terms of interest

Dickens: clustered co-occurrence graph



Dickens: clustered co-occurrence matrix



Conclusions

- Co-occurrences in a graph can be visualised using graphs and matrices
- This also makes them amenable to mathematical analysis
- Advantage: choices have to be made explicitly, e.g. what method to use, what parameter values. This does not necessarily remove choices, but at the very least shows them explicitly.
- Interesting structures can be observed

Open questions

- Which methods and parameter choices are optimal for a given context?
- Can this be extended to include a time component of corpora?
- Are there other relevant structures in corpora (besides co-occurrences) that can be represented graphically?
- Are there other structures in such graphical representations (besides densily connected 'clusters') that are linguistically relevant?