Tracing Shifting Conceptual Vocabularies Through Time

20 November 2016

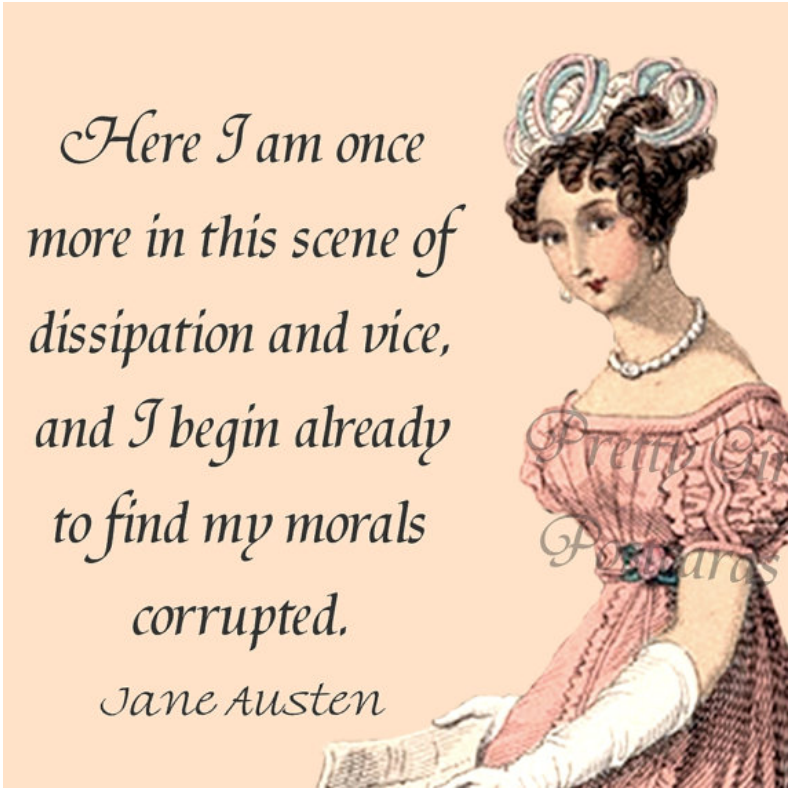
Gabriel Recchia, Ewan Jones, Paul Nulty, John Regan, & Peter de Bolla
CRASSH, The Concept Lab, University of Cambridge
GLR29@cam.ac.uk
@mesotronium

“broadcast”

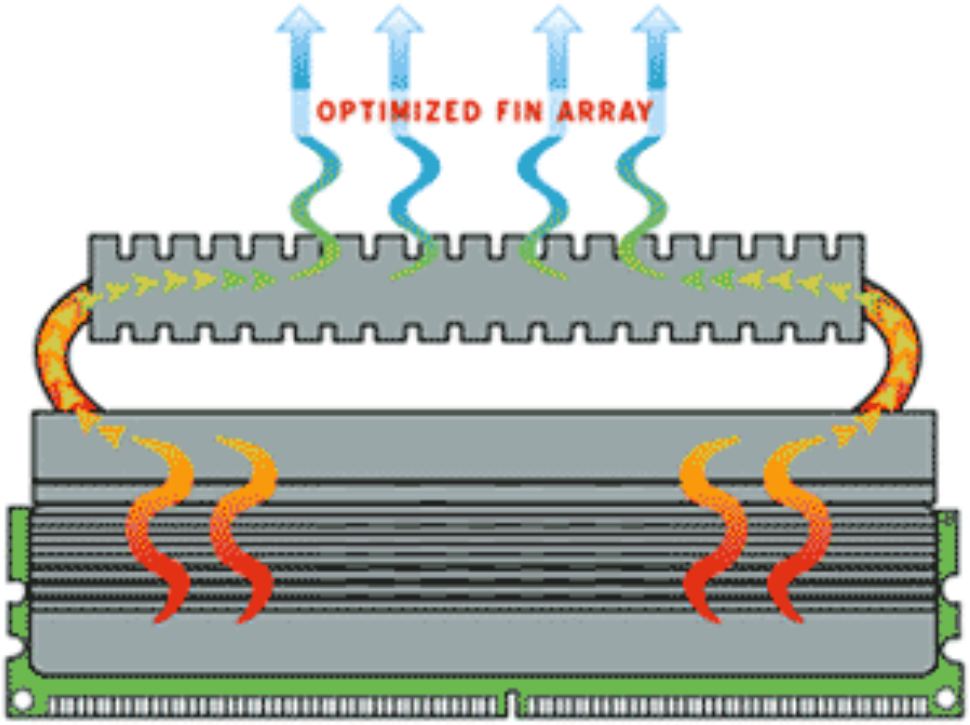
1850s

1950s

“dissipation”



1790s



2000s

1990

debauchery, extravagance, avarice, drunkenness, intemperance

1950

debauchery, dissipation, extravagance, idleness, avarice, drunkenness, intemperance, profligacy, indolence

1900

debauchery, dissipation, extravagance, idleness, profligacy, cowardice, intemperance, sensuality, indolence

1850

debauchery, dissipation, extravagance, idleness, petulance, selfishness, sloth, sensuality, gluttony

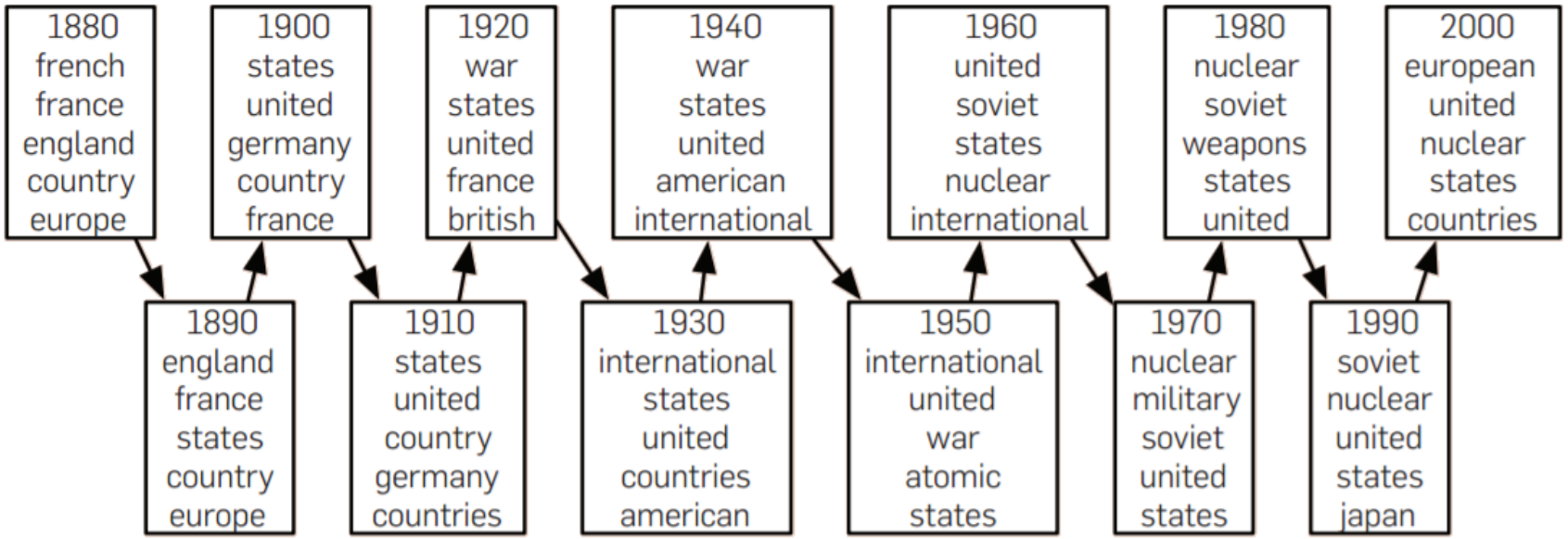
1800

debauchery, dissipation, extravagance, impertinence, laziness, selfishness, sloth, stupidity, wantonness

How do we trace the vocabulary that's associated with a particular concept over time, keeping in mind that the meanings of individual words change?

Related work

- Tracking changes in frequency of particular ‘concepts’ over time using topic models or word embeddings (Hall, Jurafsky, & Manning 2008; Wang & McCallum 2006; Blei & Lafferty 2006; Sigrist & Rawat 2009)



From Fig. 5 of 'Probabilistic Topic Models,' Blei, 2012. Communications of the ACM, 55(4), p. 81.

Related work

- Tracking changes in frequency of particular ‘concepts’ over time using topic models or word embeddings (Hall, Jurafsky, & Manning 2008; Wang & McCallum 2006; Blei & Lafferty 2006; Sigrist & Rawat 2009)

- Tracing changes in word meaning (Frermann & Lapata 2016; Mitra et al. 2015, Hamilton et al. 2016, Gulordava & Baroni 2011)

- *Concepts Through Time* (Wevers, Kenter, & Huijnen, 2015; Kenter, Wevers, & Huijnen, 2015)

bird

apple

bird

snake

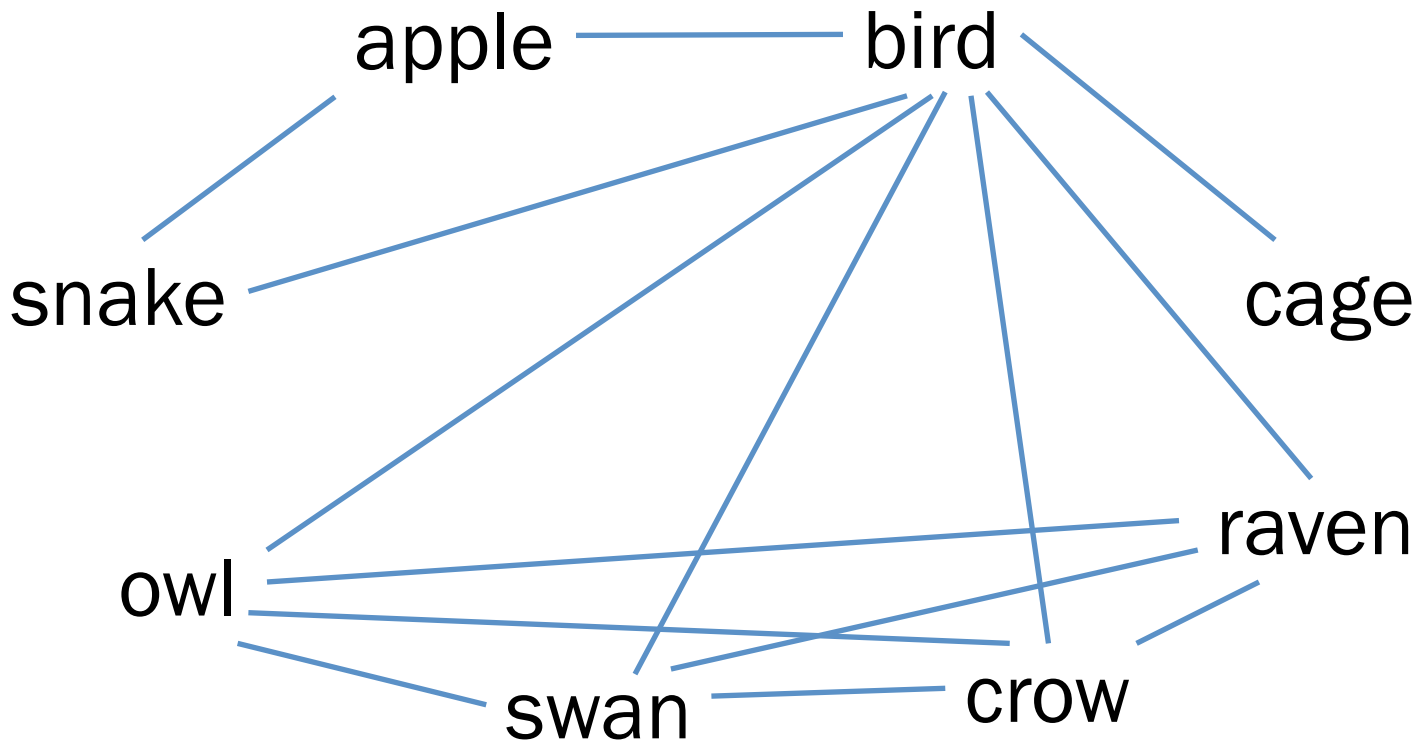
cage

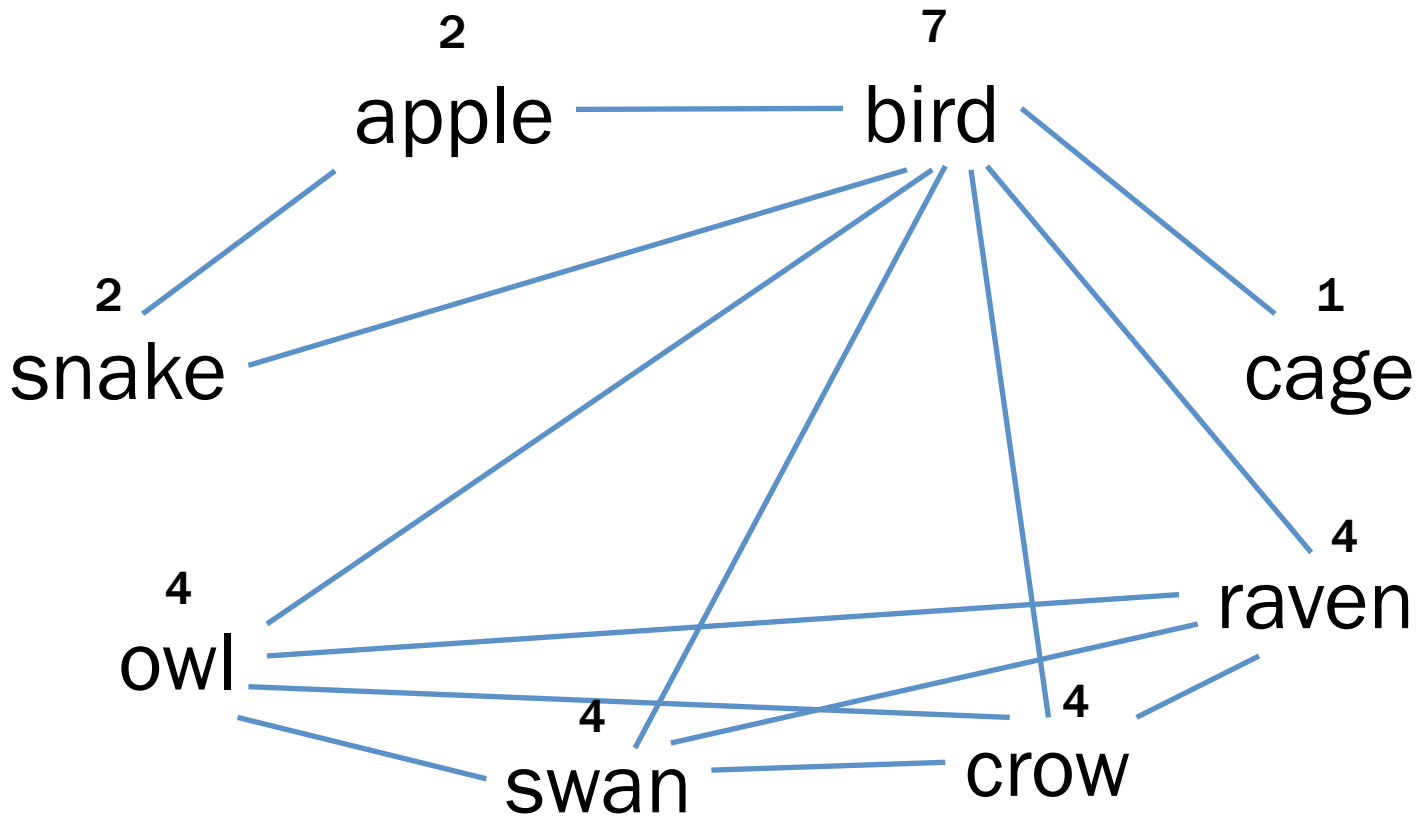
raven

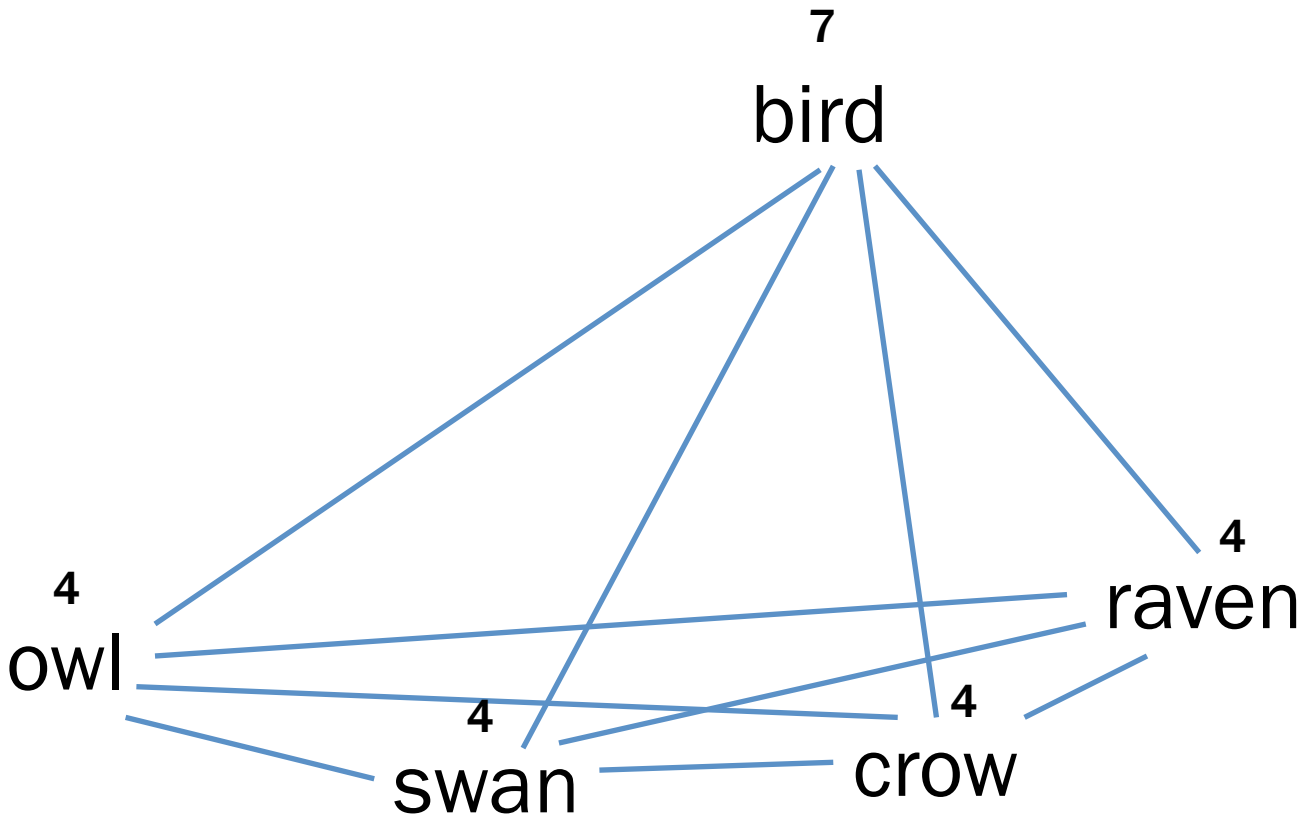
owl

swan

crow


apple
bird
snake
cage
raven
owl
swan
crow


2
apple
7
bird
2
snake
1
cage
4
raven
4
owl
4
swan
4
crow

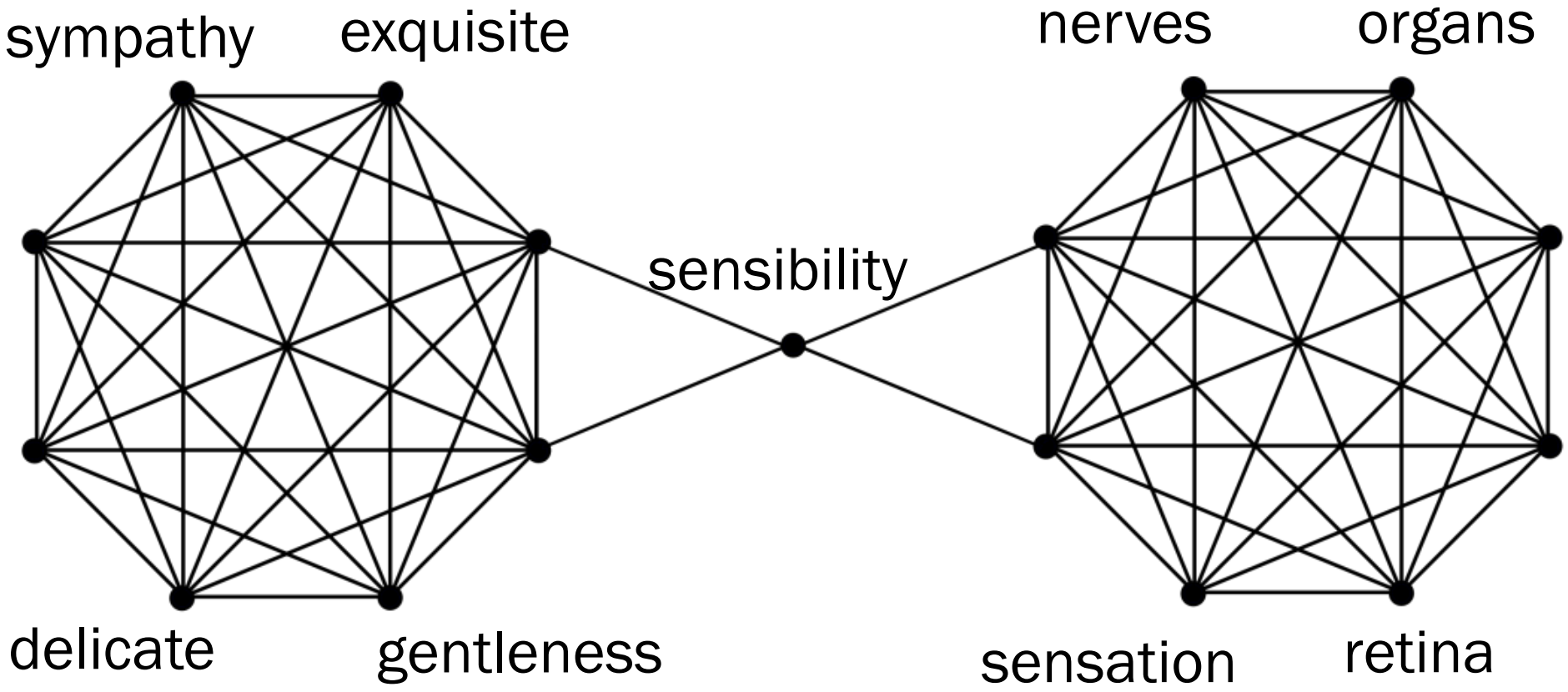

7
bird
4
raven
4
owl
4
4
swan
crow

sensibility

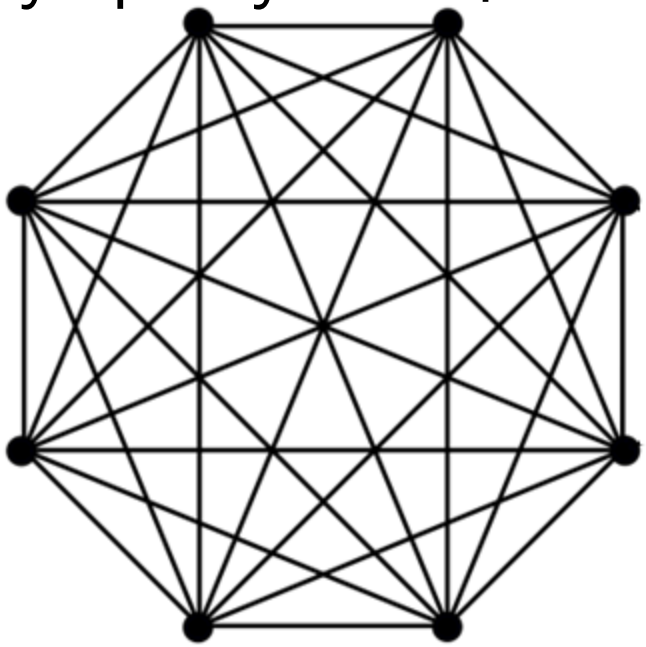
sympathy exquisite nerves organs

sensibility

delicate gentleness sensation retina

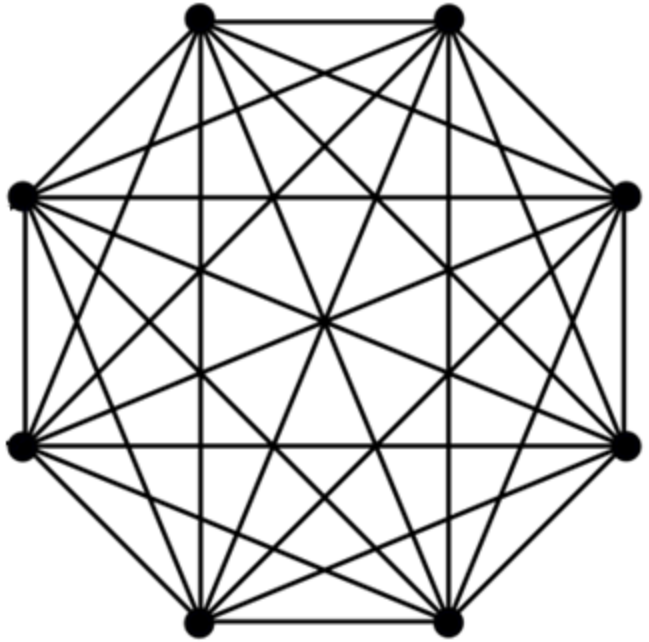

sympathy
exquisite
nerves
organs
sensibility
delicate
gentleness
sensation
retina

sympathy exquisite



delicate gentleness

nerves organs



sensation retina

Our method

- Subnetwork is considered a “conceptual network” only if *all* words in the network are highly related to *all* other words in the network
  - *e.g.,* network is a k-clique after all edges not meeting some weight threshold have been removed

- For the purposes of this talk:
  - *Nodes* represent *words*
  - *Weighted edges* represent similarity/relatedness relations, as quantified by applying cosine similarity to the *Histwords* dataset of Hamilton, Leskovec & Jurafsky (English only, SGNS word2vec vectors)

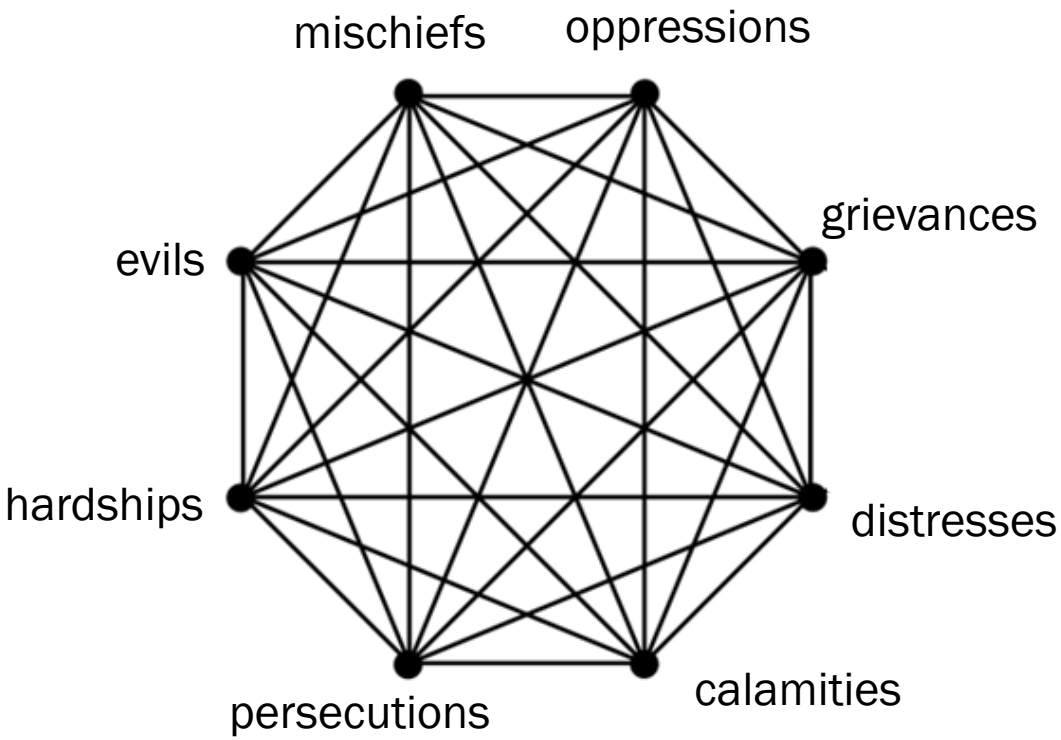
Our method

- Given a size *k* and a set of seed words *W…*
  - k = 9
  - W = { “grievances” }

…find the fully connected graph of size *k* containing all words in W such that the minimum edge weight is as high as possible

Our method

- Given a size *k* and a set of seed words *W…*
  - k = 8
  - W = { “grievances” }



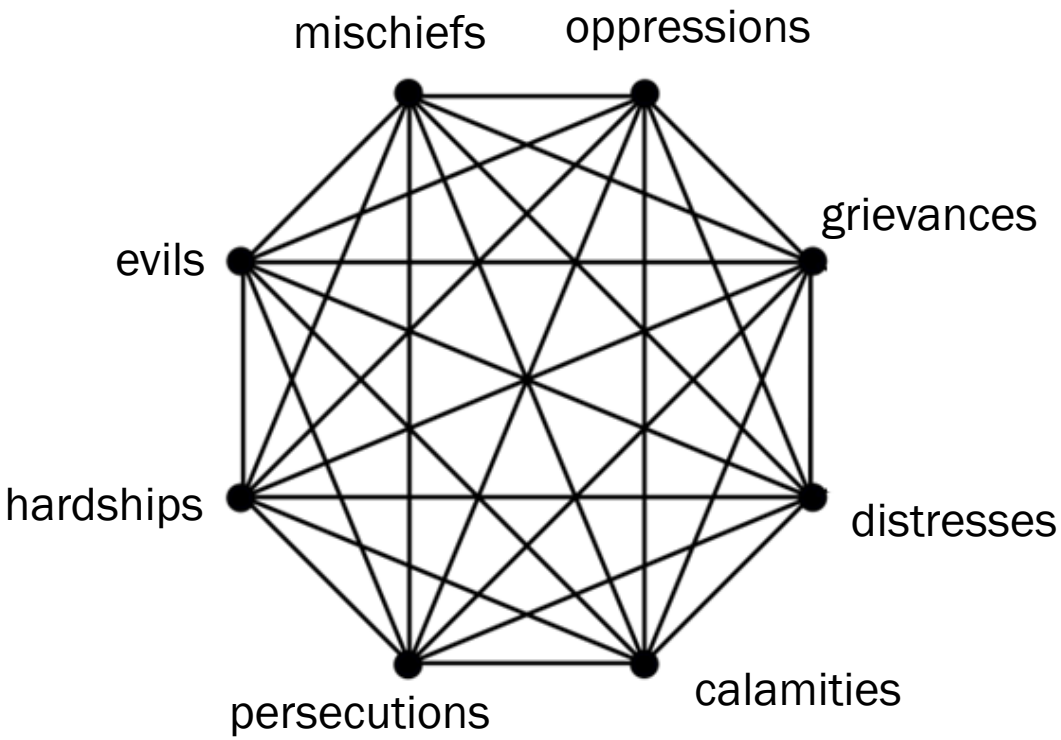
Our method

- Updating from decade to decade: the "drop one, add one" rule

  – "Is it possible to increase the minimum edge weight by replacing one of these nodes with a node currently not in the subgraph? If so, which of all possible replacements would increase the minimum edge weight the most?"

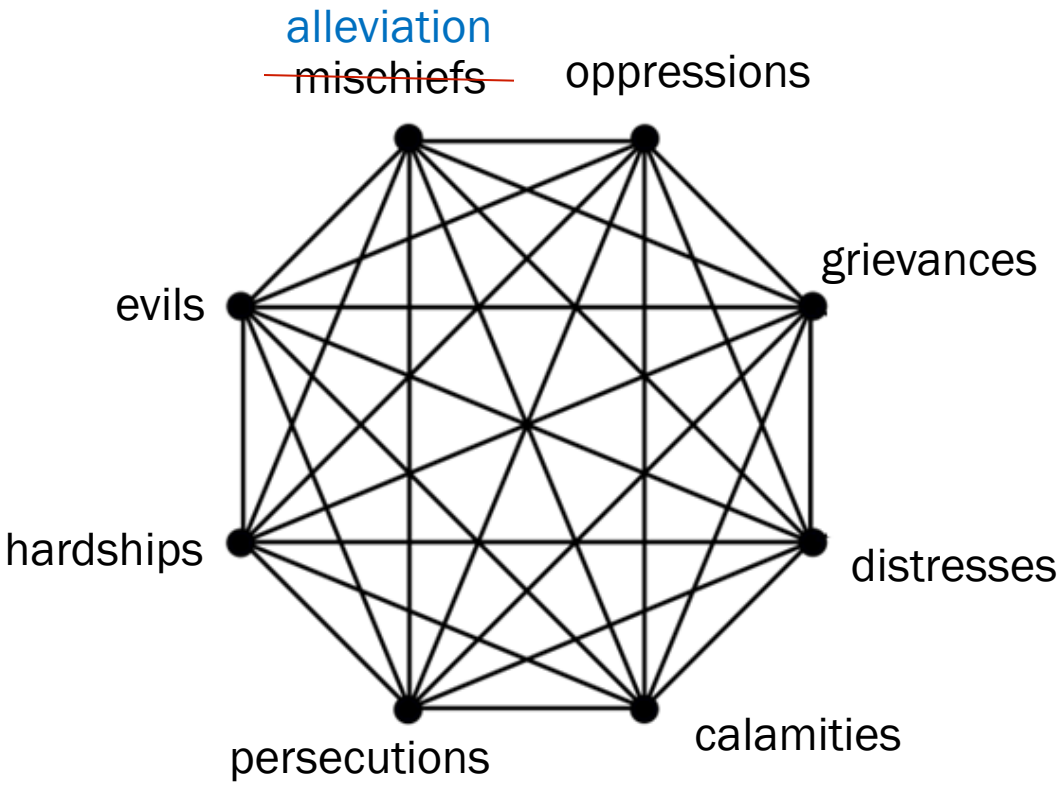
Our method

- Given a size *k* and a set of seed words *W*…
  - k = 8
  - W = { “grievances” }



Our method

- Given a size *k* and a set of seed words *W*…
  - k = 8
  - W = { “grievances” }



1800	calamities,distresses,evils,grievances,hardships,mischiefs,miseries,oppressions,persecutions
1810	calamities,distresses,evils,grievances,hardships,alleviation,miseries,oppressions,persecutions
1820	calamities,distresses,evils,grievances,hardships,alleviation,miseries,oppressions,burthens
1830	calamities,distresses,evils,grievances,hardships,alleviation,miseries,alleviate,burthens
1840	calamities,distresses,evils,grievances,hardships,alleviation,miseries,alleviate,sufferings
1850	calamities,distresses,evils,grievances,hardships,privations,miseries,alleviate,sufferings
1860	calamities,distresses,evils,grievances,hardships,privations,miseries,vexations,sufferings
1870	calamities,distresses,evils,grievances,hardships,privations,miseries,vexations,misfortunes
1880	calamities,distresses,evils,grievances,ills,privations,miseries,vexations,misfortunes
1890	calamities,distresses,evils,grievances,dangers,privations,miseries,vexations,misfortunes
1900	calamities,distresses,evils,grievances,dangers,privations,miseries,inconveniences,misfortunes
1910	calamities,distresses,evils,grievances,dangers,privations,miseries,inconveniences,hardships
1920	calamities,distresses,evils,anxieties,dangers,privations,miseries,inconveniences,hardships
1930	calamities,distresses,sufferings,anxieties,dangers,privations,miseries,inconveniences,hardships
1940	calamities,distresses,sufferings,misfortunes,dangers,privations,miseries,inconveniences,hardships
1950	calamities,distresses,sufferings,misfortunes,dangers,privations,miseries,perils,hardships
1960	calamities,distresses,sufferings,misfortunes,discouragements,privations,miseries,perils,hardships

Basic evaluation

- *Flexibility:* Does the network allow words to freely drop in and out? How frequently does this happen for the seed word(s)?

- *Stability:* Does this network have a core contingent that stays somewhat constant over time, or is it changing just as much as it would have if we just randomly chose a word to replace every timestep?

Basic evaluation

- *Flexibility:* the seed word used to generate the initial size-9 network in 1800 was no longer present in the 1990 network in 147 of 212 cases (69%)

- *Stability:* average overlap in vocabulary between the initial 1800s network and the final 1990s network was 33%

Basic evaluation

Even when vocabulary changes, concept generally remains similar…

1800: *anxieties, dejected, dejection, distraction, fits, insupportable, languishing, uneasy, weariness*

1990: *anxieties, grief, despair, disappointment, misery, sorrow, anguish, sadness, loneliness*

Basic evaluation

Even when vocabulary changes, concept generally remains similar…

1800: *battery, bullet, cannon, flanked, musket, muskets, pikes, pounders, rods*

1990: *battery, batteries, cannon, gun, howitzers, rifles, rifle, mortars, guns*

Basic evaluation

…albeit for some words less so than others

1800: *abstruse, definitions, disquisition, disquisitions, explanations, explication, grammatical, illustrating, logical*

1990: *abstruse, mathematical, philosophy, theory, metaphysics, metaphysical, empirical, theoretical, philosophical*

- Networks available: http://nowin2d.com/vocabularies/

Towards a ‘real’ evaluation

BRITISH LIBRARY

COLLECTION METADATA
Data Services

Home > Collection Metadata > Data Services > Downloads

Downloads

British National Bibliography

Books and serials eligible for BNB. Models, schema and URI patterns available here. Updated monthly. Zipped folders include multiple files and a PDF document.

DATASET	DATE	SIZE (KB) *(full file)*	FULL FILE	SAMPLE
BNB LOD Books	2016-10	1,100,945	nt	nt
BNB LOD Serials	2016-10	42,470	nt	nt
BNB LOD Books	2016-10	1,125,001	rdfxml	rdfxml
BNB LOD Serials	2016-10	37,170	rdfxml	rdfxml
VoID Descriptions	2016-10	5	ttl	N/A

“Journal of ‘X’”

<http://bnb.data.bl.uk/id/concept/lcsh/Psychiatry>
1876 : nervous and mental disease

<http://bnb.data.bl.uk/id/concept/lcsh/Engineering>
1921 : applied mathematics and mechanics

<http://bnb.data.bl.uk/id/concept/lcsh/Entrepreneurship>
1985 : business venturing

<http://bnb.data.bl.uk/id/concept/lcsh/Tourism>
1972 : travel research

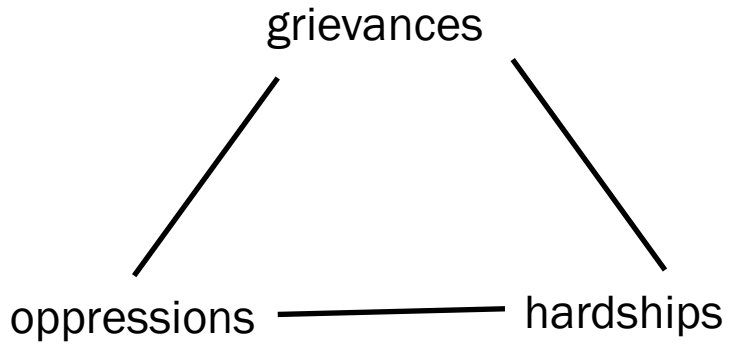
Future work

- Optimize initialization parameters
- Apply relevant ideas from the field of ontology evolution (Pesquita & Couto, 2012; Cano-Basave, Osborne & Salatino, 2016; Wang et al., 2105)
- Create ground truth dataset

Thank You

Our method

- Given a size *k* and a set of seed words *W…*
  - k = 8
  - W = { “grievances” }



Our method

- Given a size *k* and a set of seed words *W…*
  - k = 8
  - W = { “grievances” }