

Exploring Author Influence with Networks and NLP

Benjamin Ortiz Ulloa

10/27/2018

What We'll Cover

1. **Graphs: An Introduction**
2. **Graphs: Collaboration Networks**
3. **NLP: An Introduction**
4. **NLP: Topic Modelling**

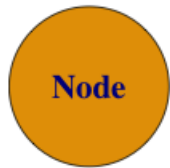
Graphs are made up of things



.

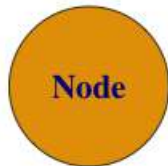
.

Things are also known as nodes

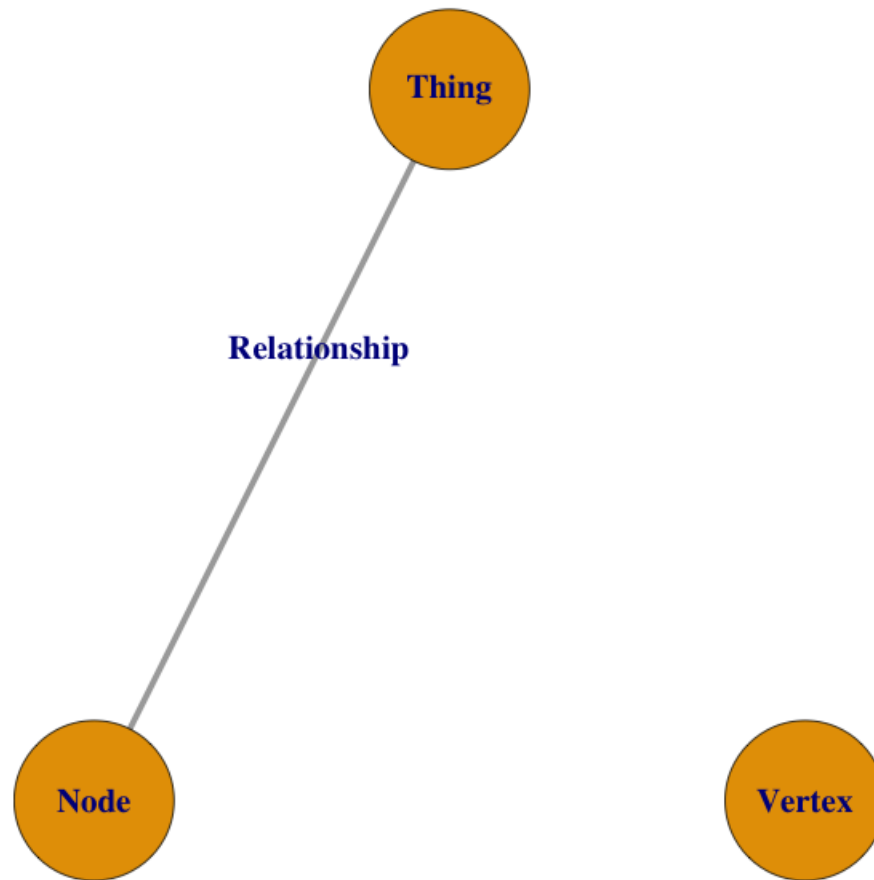


.

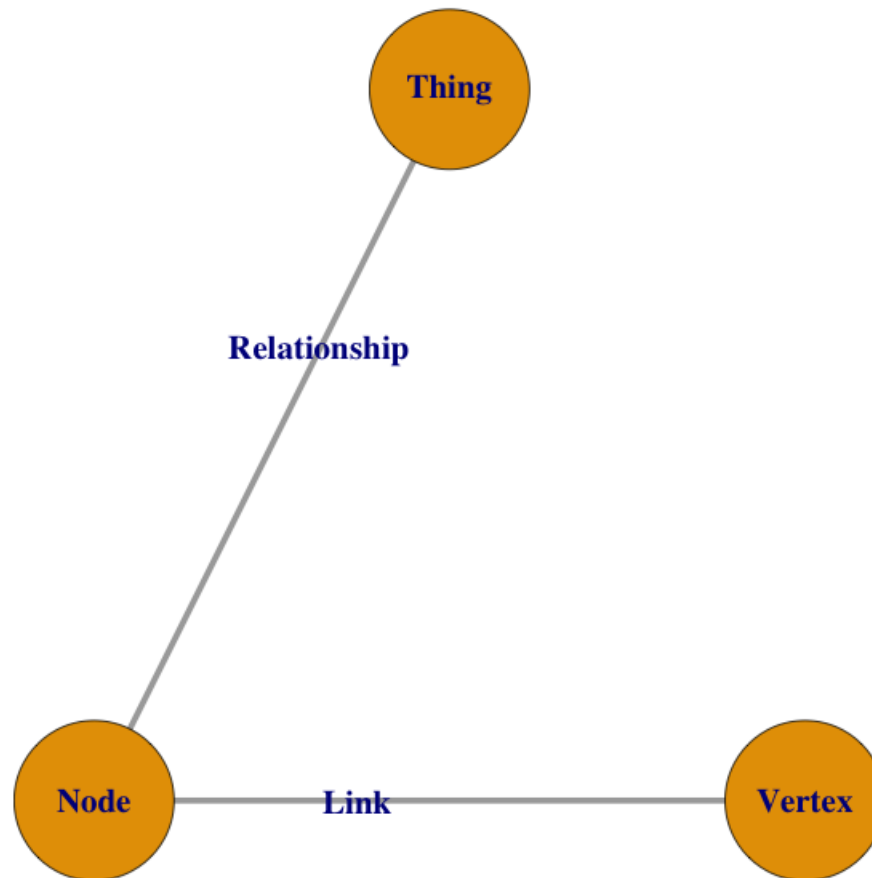
Things are also known as vertices



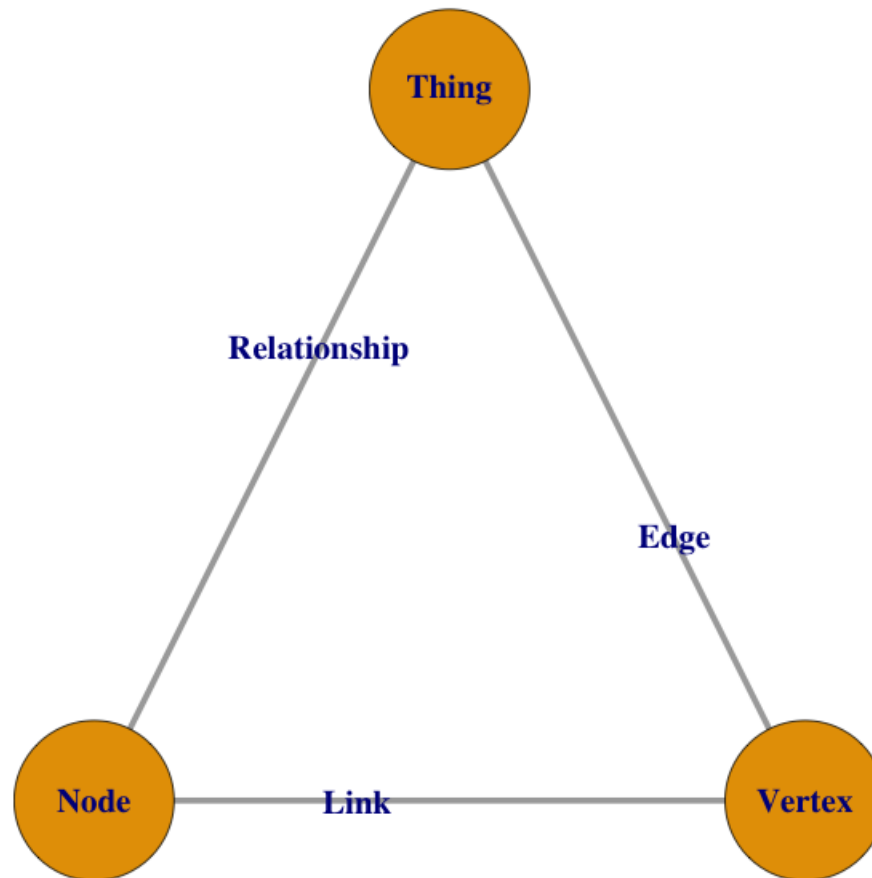
Things are connected by relationships



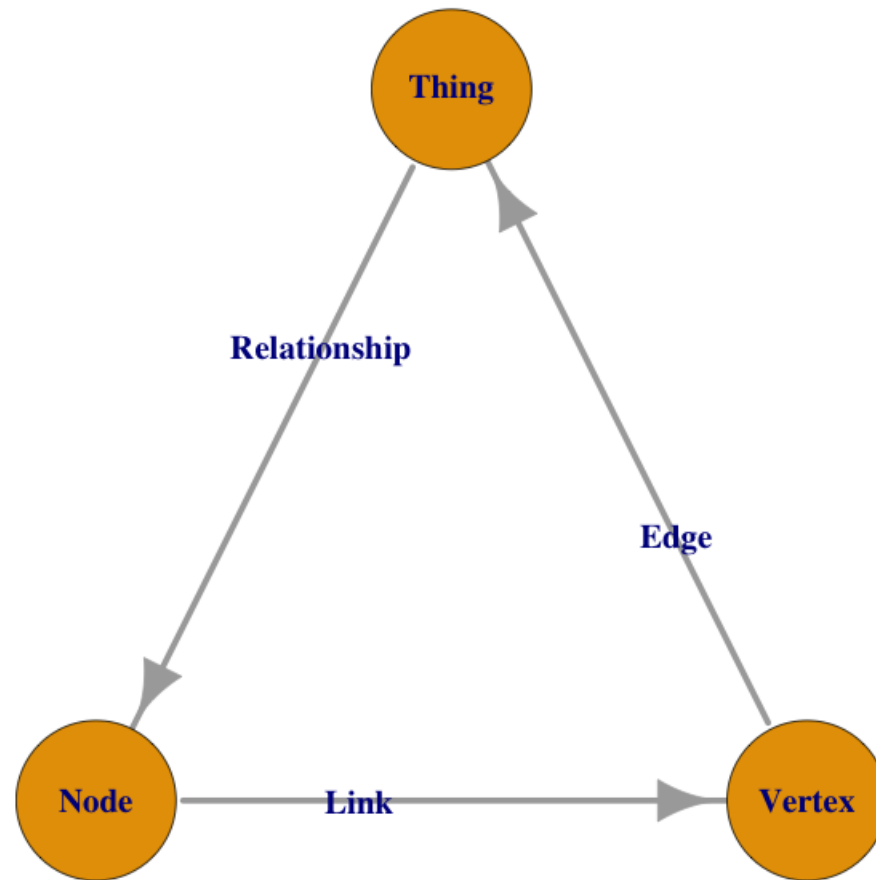
Relationships are also known as links



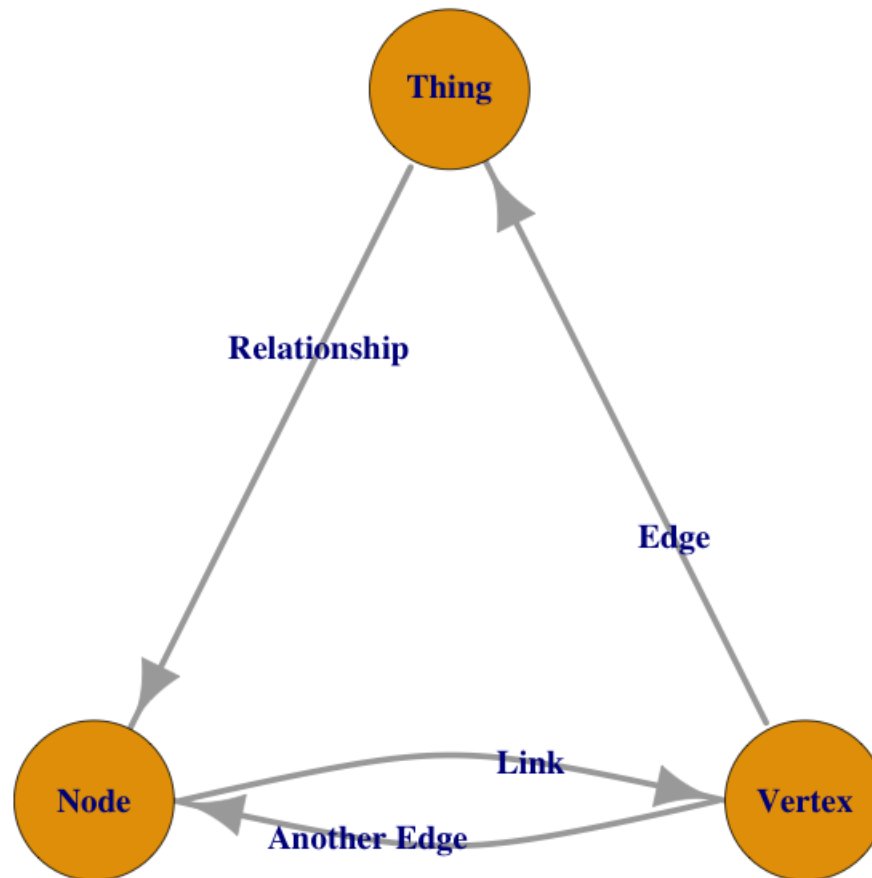
Relationships are also known as edges



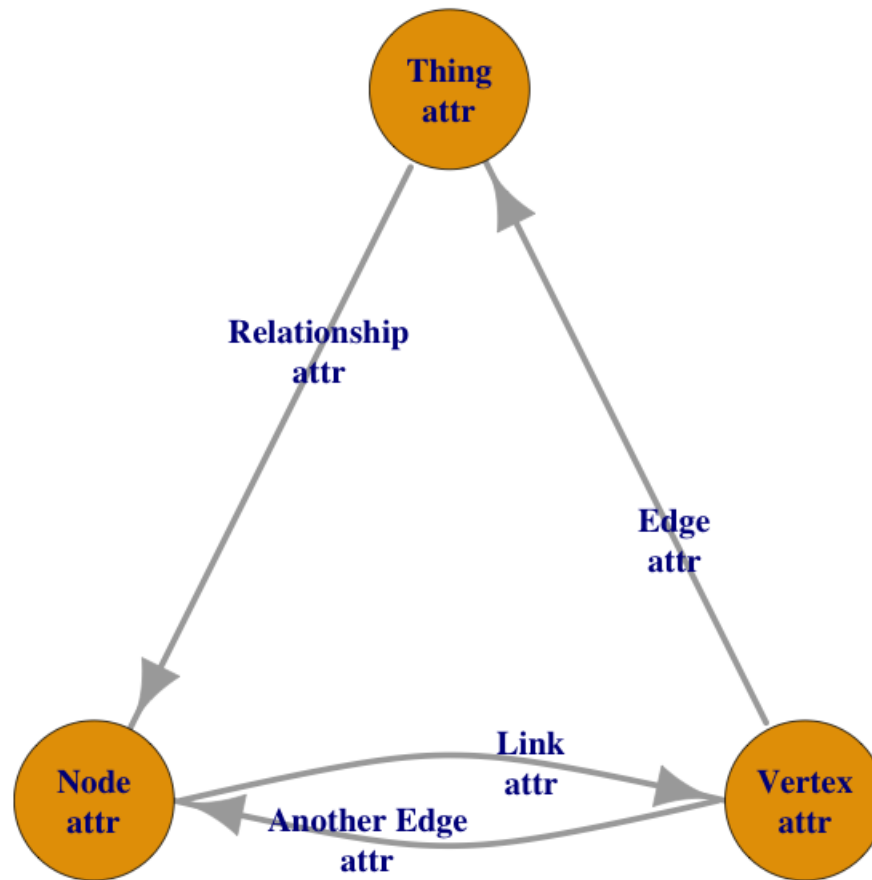
Relationships can have direction



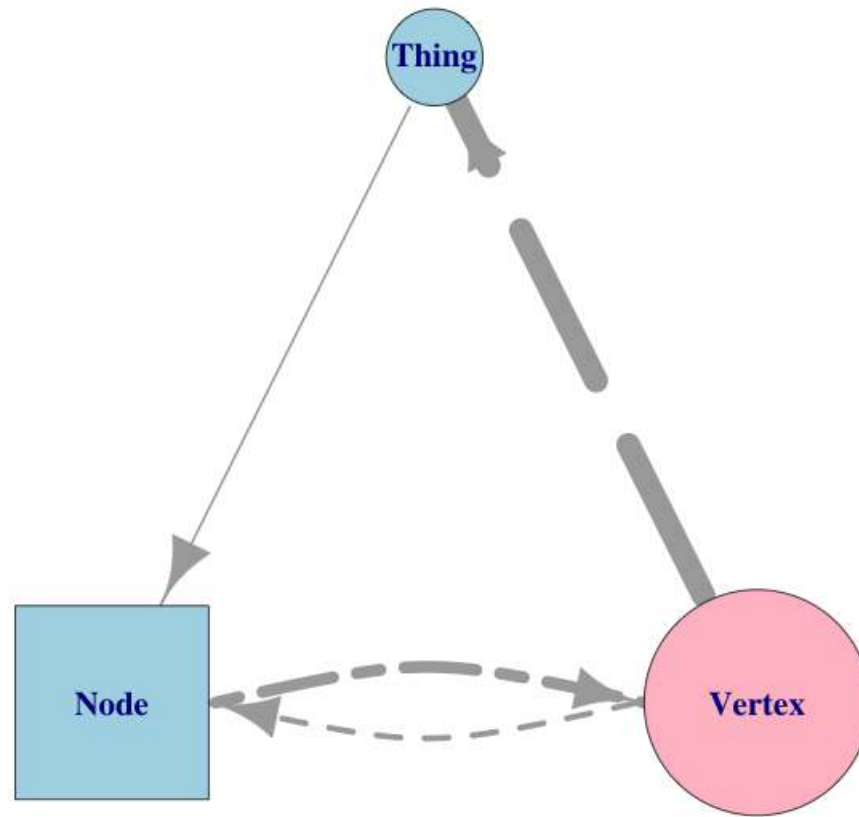
Things can have multiple relationships



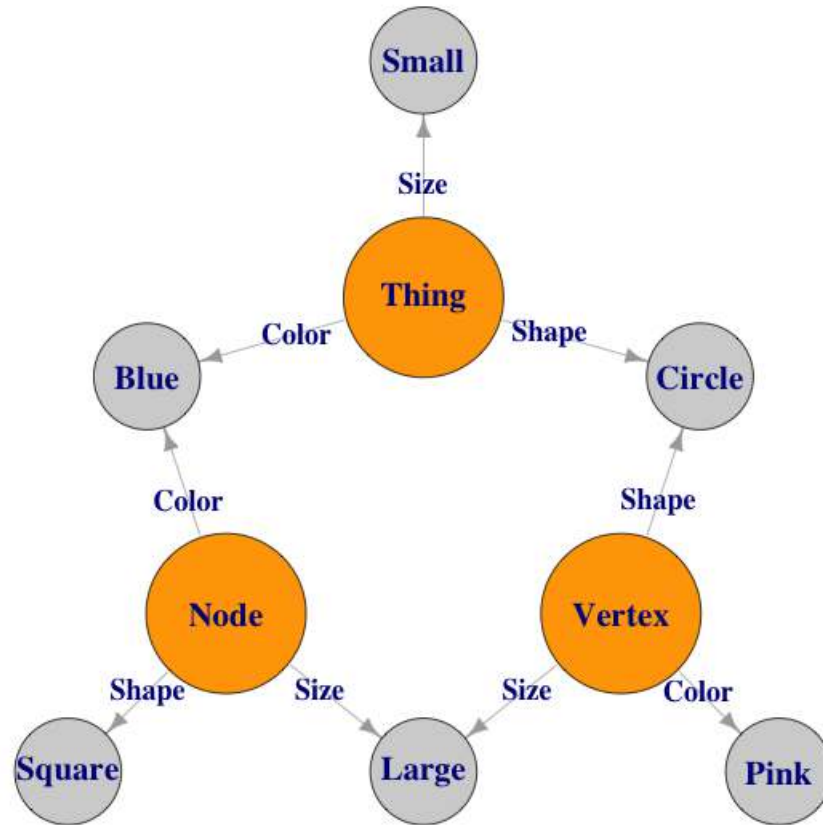
Everything can have attributes



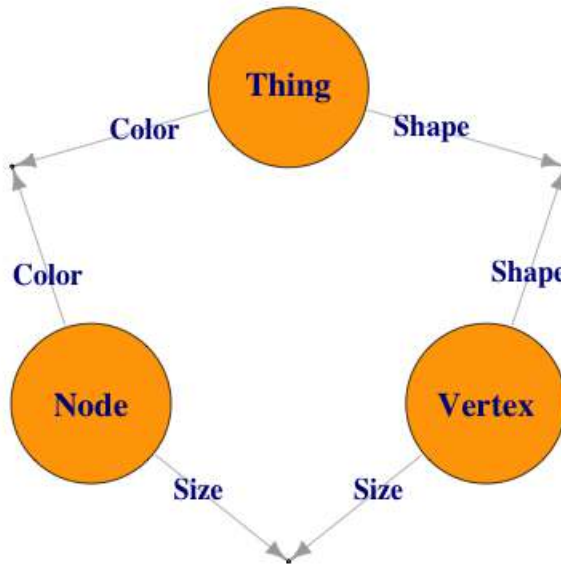
Attributes can be visualized



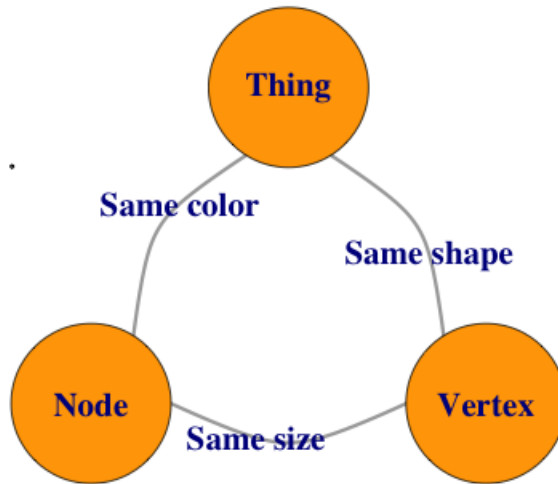
Node attributes can be represented as nodes themselves



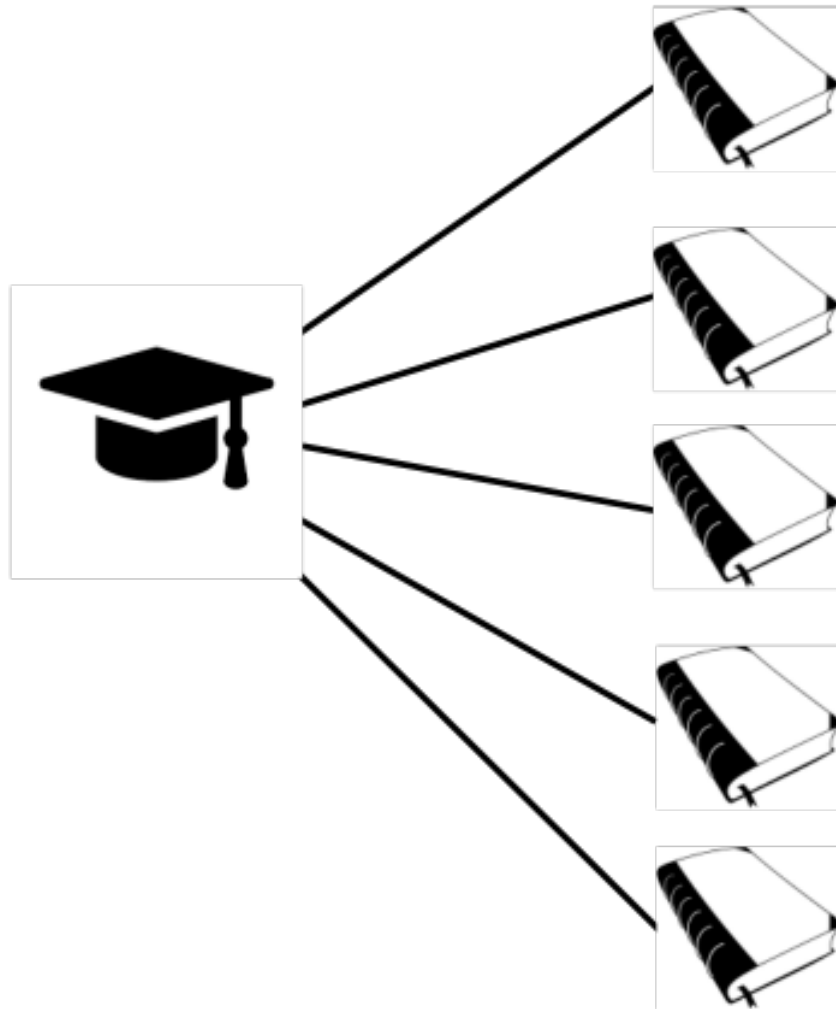
Direct relationships can be inferred from indirect relationships



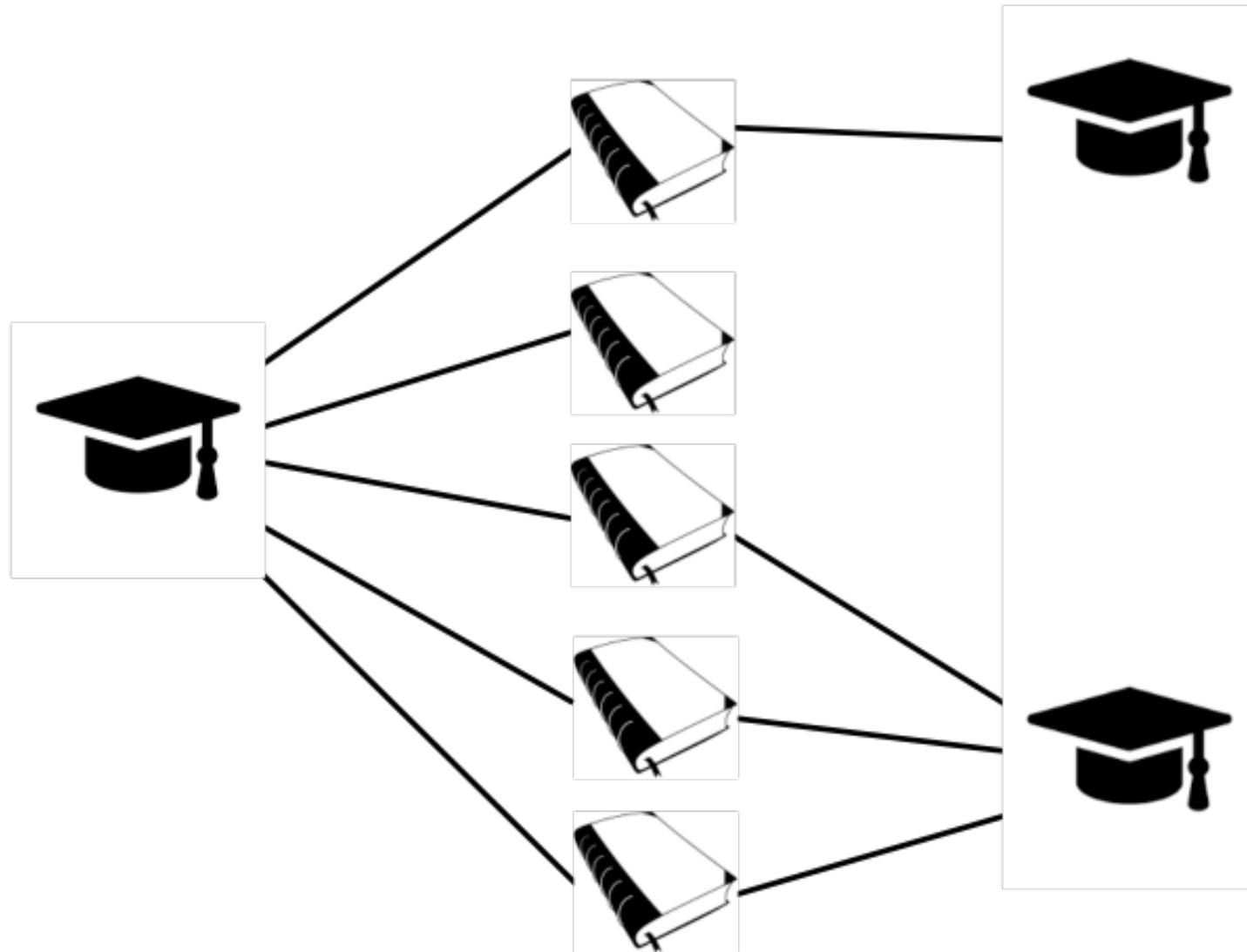
These projected relationships enrich our data



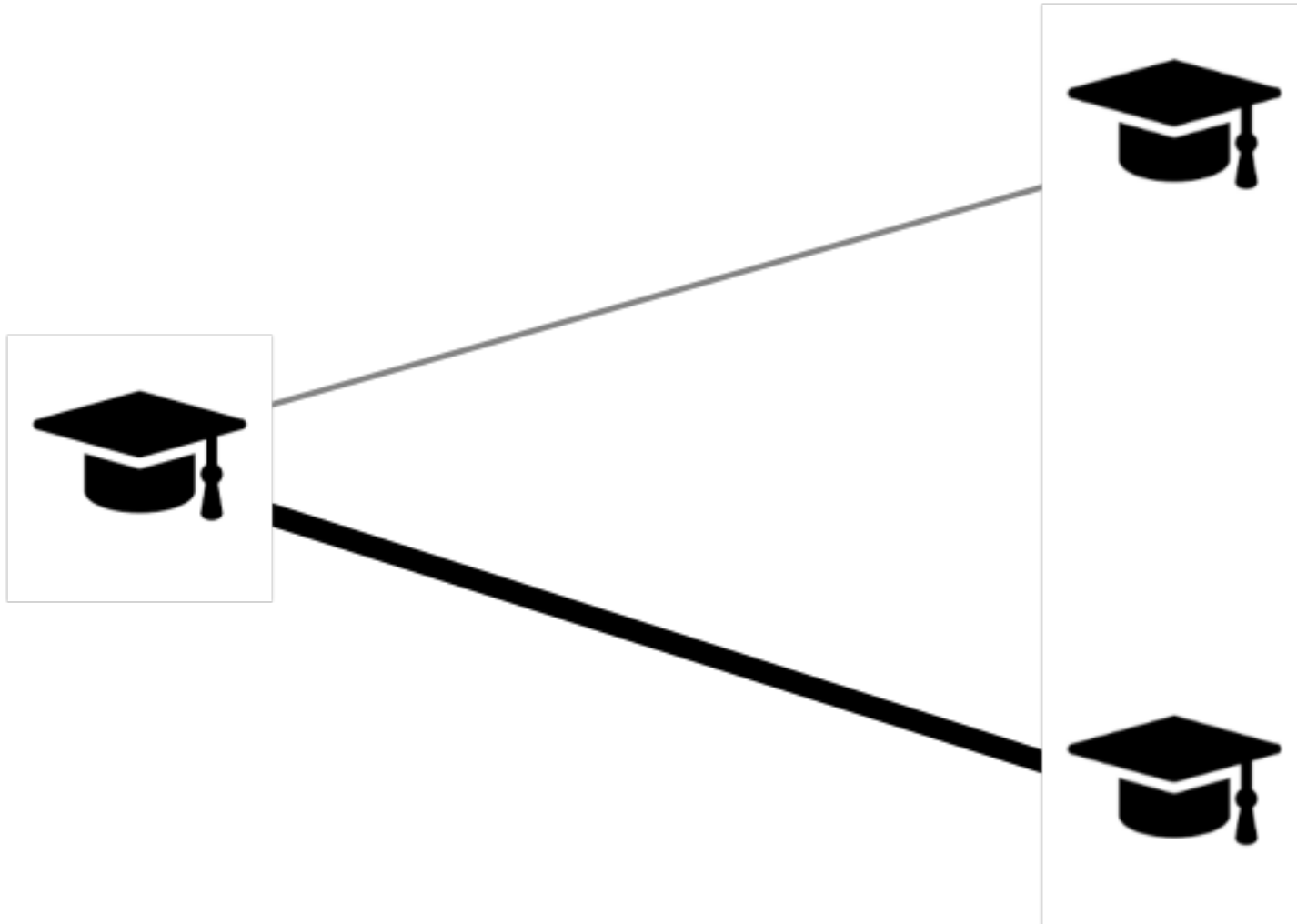
An example of a bipartite graph: an author writes publications



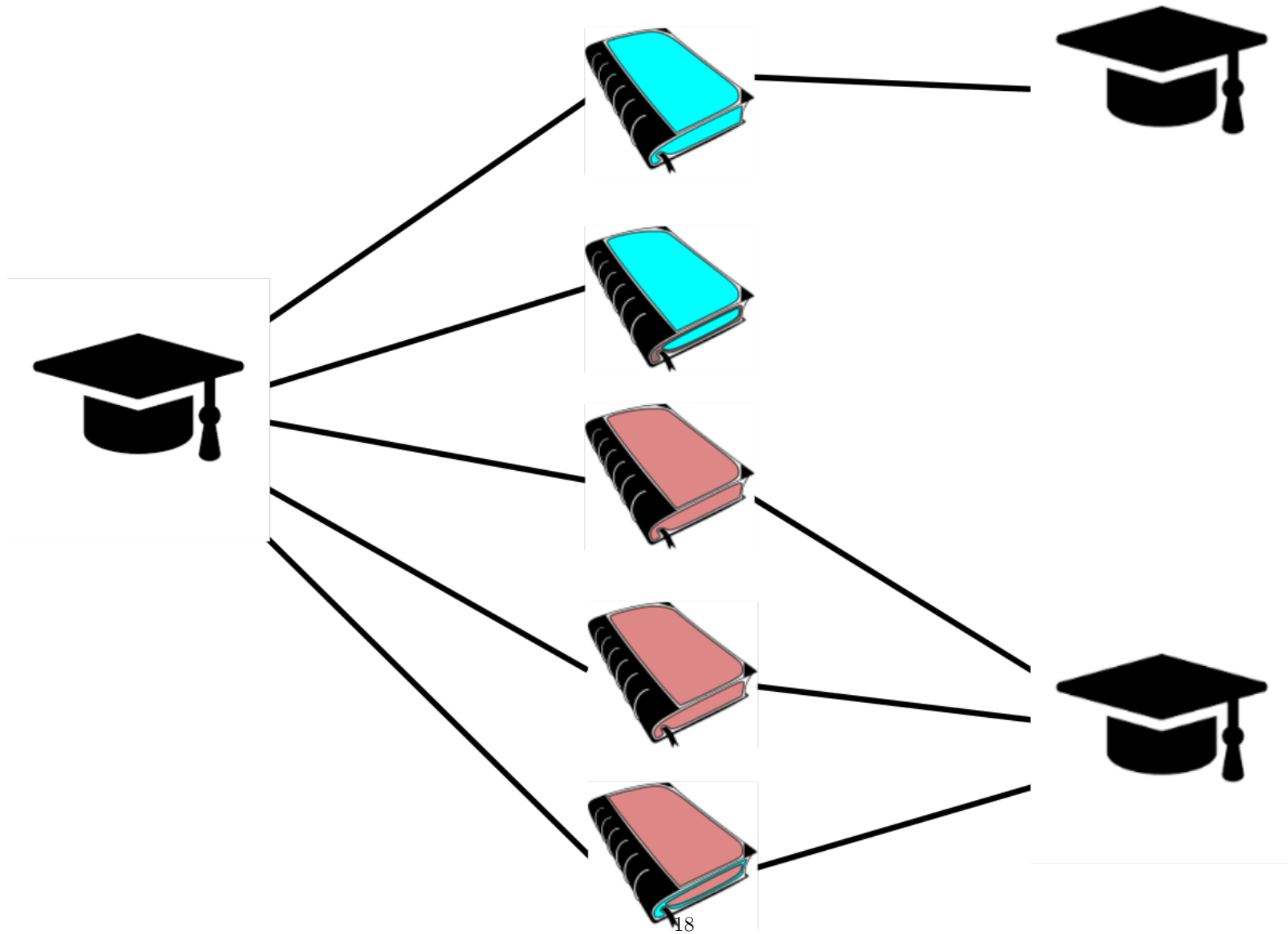
Other authors collaborated on those publications



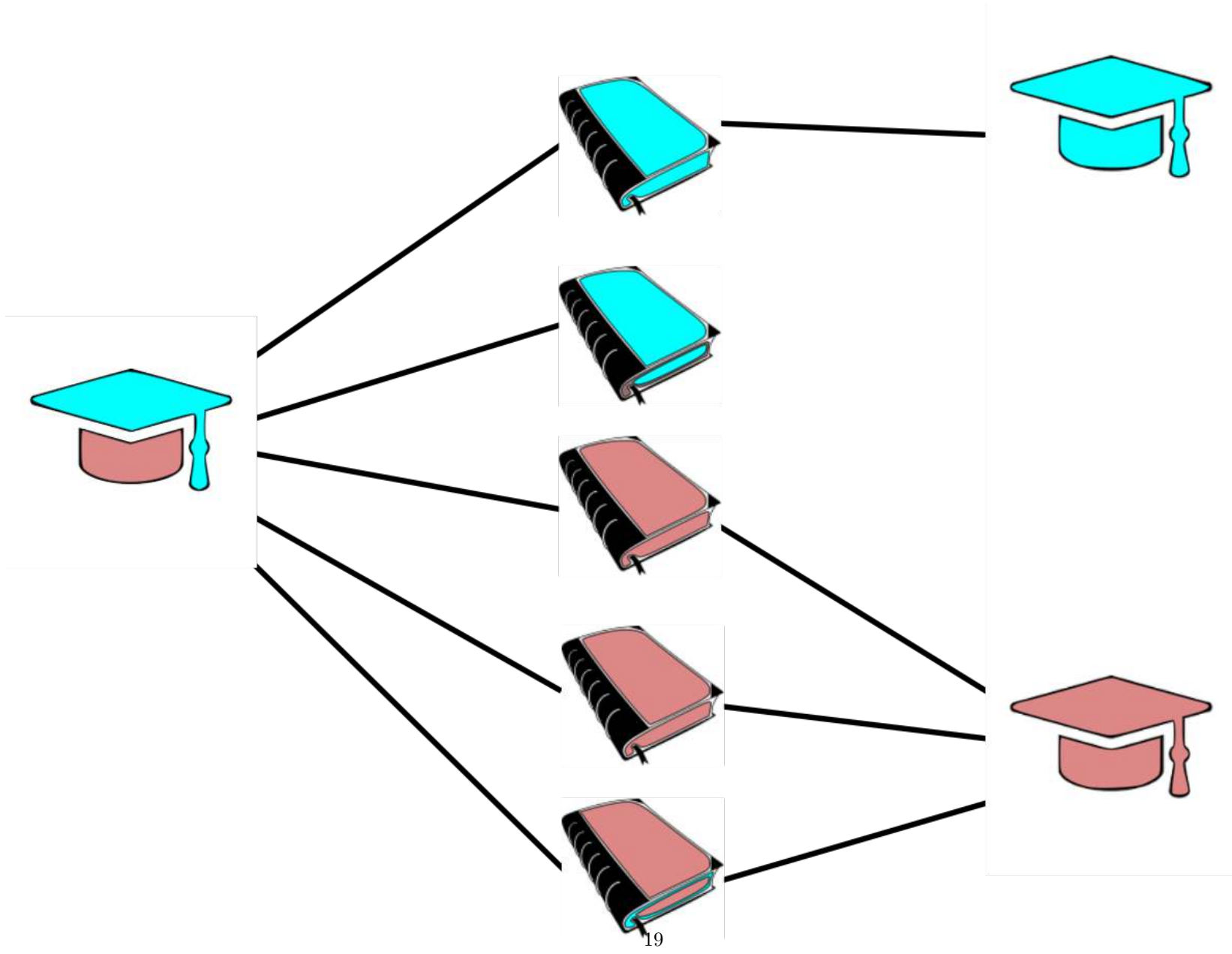
We can now infer a relationship



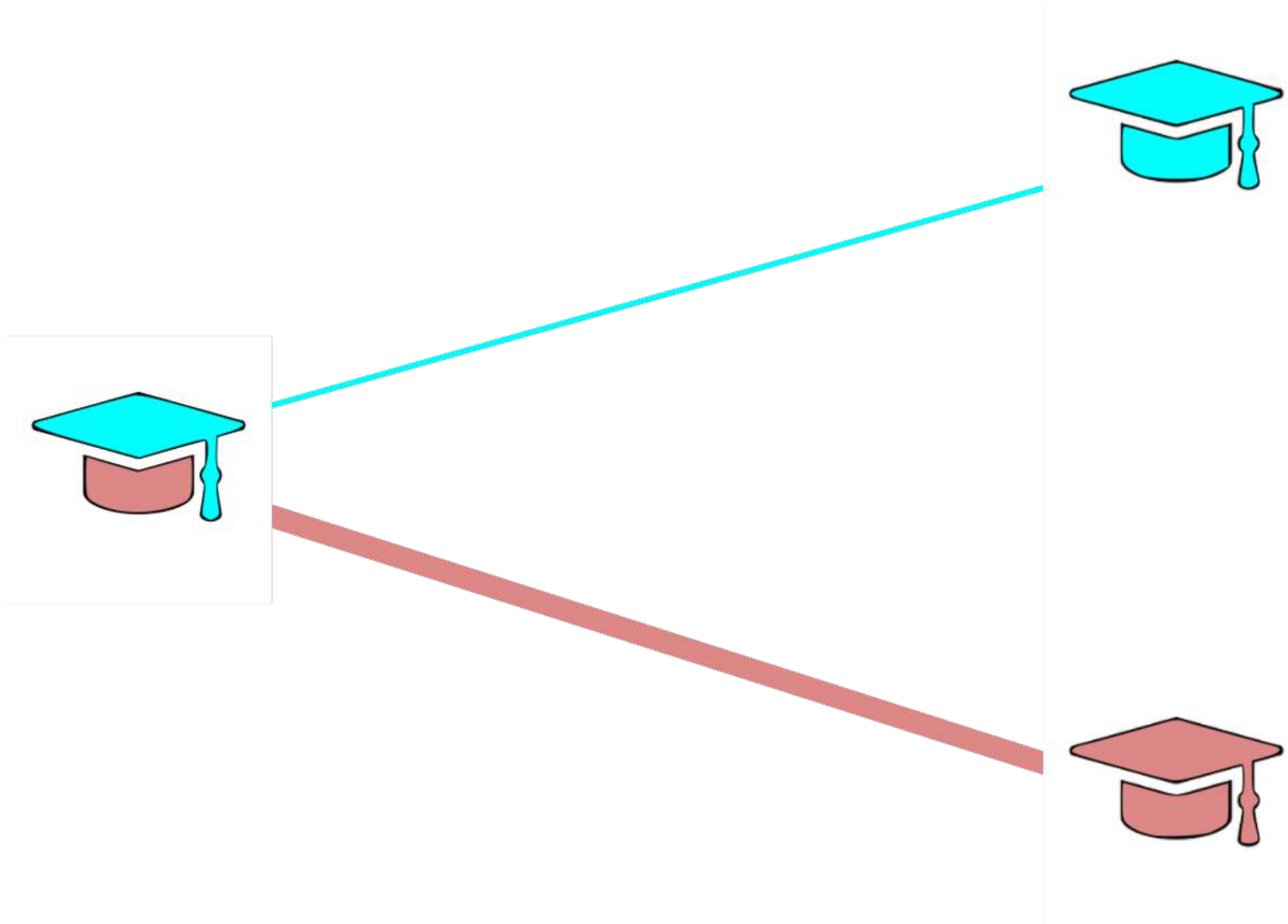
Information can be stored in publications



Information can be transferred to authors

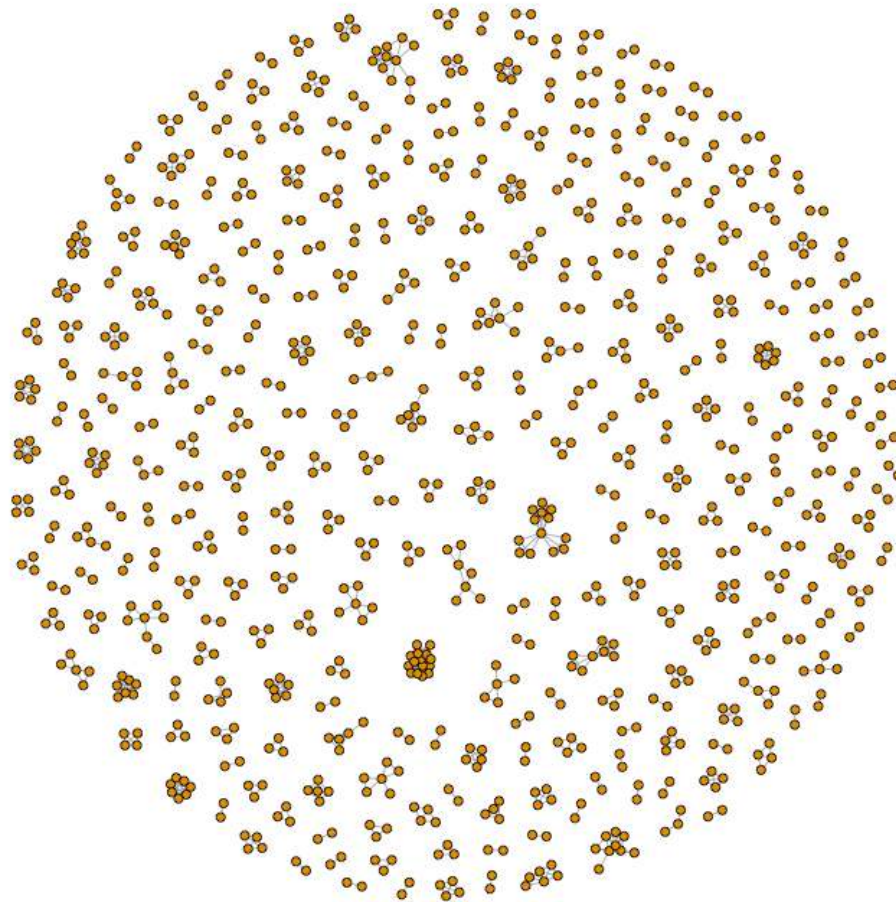


Information can be stored in relationships

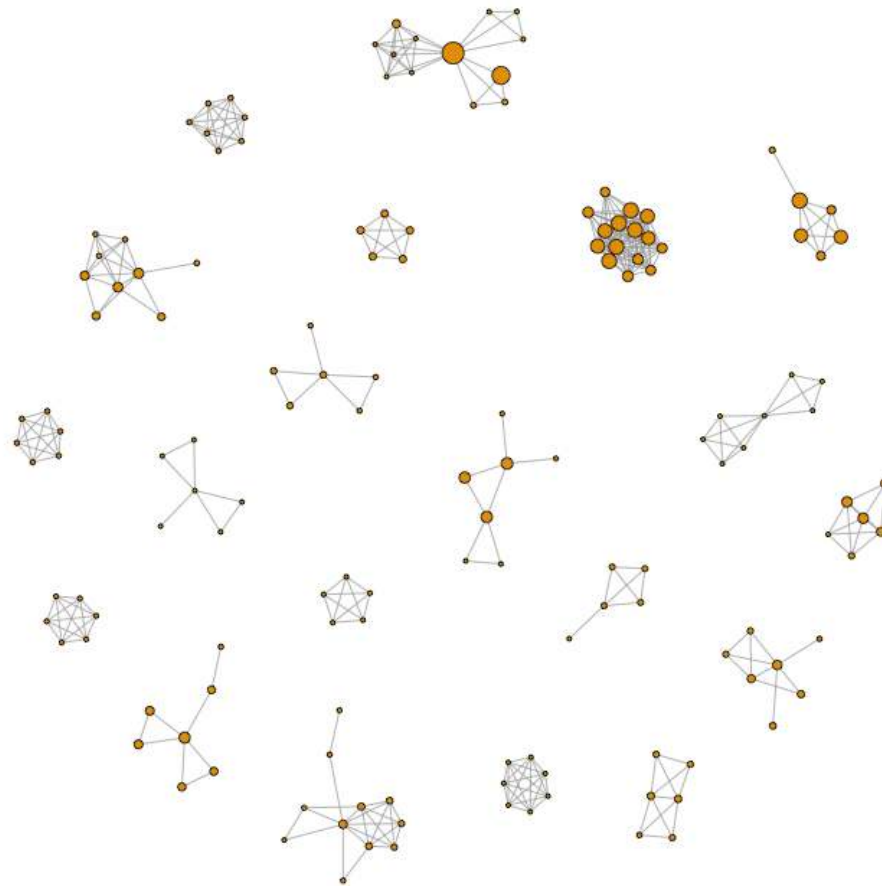


Collaboration Network:

Authors writing about network science & agent based models



20 largest components for exploration



Bag of Words

The fox jumped over the other fox

The dog ran at the foxes

The foxes ran away from the dog

The cat napped on my lap

The lion napped on the hill

The cat jumped on the fox

A standard data structure for text analysis is the Document Term Matrix (DTM). This is a matrix in which the rows represent the documents in your corpus and the columns represent every word

	at	away	cat	dog	fox	foxes	from	hill	jumped	lap	lion	my	napped	on	other	over	ran	the
doc_1	0	0	0	0	2	0	0	0	1	0	0	0	0	0	1	1	0	2
doc_2	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	2
doc_3	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	1	1
doc_4	0	0	1	0	0	0	0	0	0	1	0	1	1	1	0	0	0	1
doc_5	0	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	0	2
doc_6	0	0	1	0	1	0	0	0	1	0	0	0	0	1	0	0	0	2

Remove Stop Words

For computational reasons, we generally want to remove words that are so common they don't really provide us with significant meaning or context.

The fox jumped over the other fox

The dog ran at the foxes

The foxes ran away from the dog

The cat napped on my lap

The lion napped on the hill

The cat jumped on the fox

	cat	dog	fox	foxes	hill	jumped	lap	lion	napped	ran
doc_1	0	0	2	0	0	1	0	0	0	0
doc_2	0	1	0	1	0	0	0	0	0	1
doc_3	0	1	0	1	0	0	0	0	0	1
doc_4	1	0	0	0	0	0	1	0	1	0
doc_5	0	0	0	0	1	0	0	1	1	0
doc_6	1	0	1	0	0	1	0	0	0	0

Unsupervised Topic Modelling

Latent Dirichlet Allocation (LDA) and Structural Topic Modelling (STM) are two methods of classifying text without providing clues to what we want. These methods are great if we have a large corpus of documents, but we don't really know what we're looking for. The only parameter needed is the number of categories we want returned

Unsupervised Topic Modelling: $K = 3$

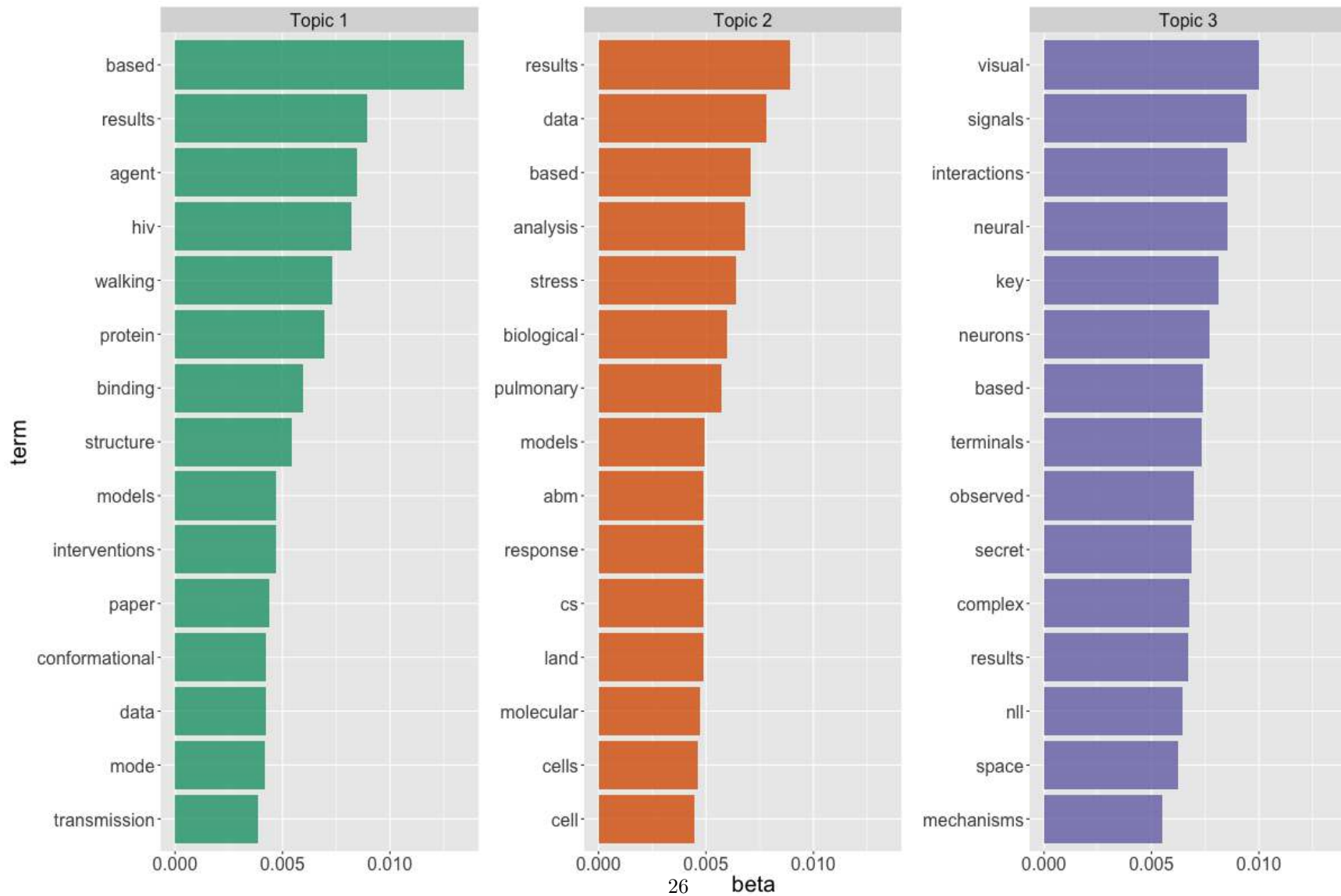
	cat	dog	fox	foxes	hill	jumped	lap	lion	napped	ran
doc_1	0	0	2	0	0	1	0	0	0	0
doc_2	0	1	0	1	0	0	0	0	0	1
doc_3	0	1	0	1	0	0	0	0	0	1
doc_4	1	0	0	0	0	0	1	0	1	0
doc_5	0	0	0	0	1	0	0	1	1	0
doc_6	1	0	1	0	0	1	0	0	0	0

Unsupervised Topic Modelling: $K = 2$

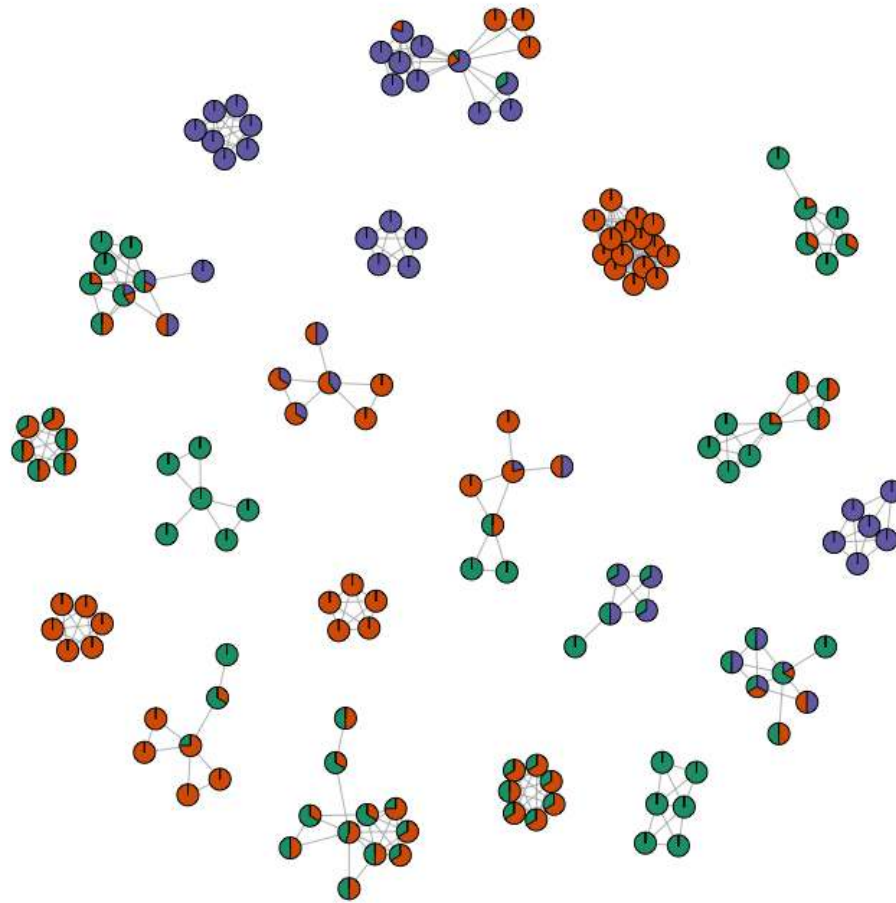
	cat	dog	fox	foxes	hill	jumped	lap	lion	napped	ran
doc_1	0	0	2	0	0	1	0	0	0	0
doc_2	0	1	0	1	0	0	0	0	0	1
doc_3	0	1	0	1	0	0	0	0	0	1
doc_4	1	0	0	0	0	0	1	0	1	0
doc_5	0	0	0	0	1	0	0	1	1	0
doc_6	1	0	1	0	0	1	0	0	0	0

STM Categories: $K = 3$

If we run an STM model on all the publications in our collaboration network, we get the following three categories

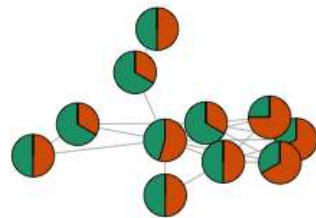
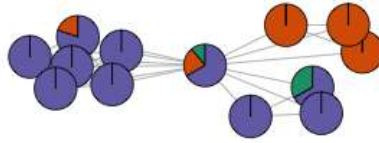


20 Largest Components: STM Categories



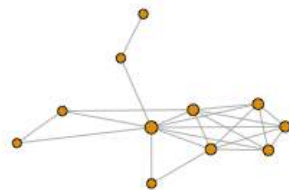
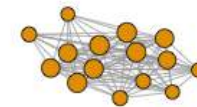
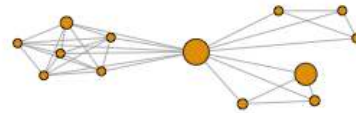
But why are we running an unsupervised algorithm?

What if a subject matter expert sees three clusters of authors and says, “These authors come from very focused labs and work on very different things.”?



Start Over

We can use that information to define the categories we want. We have a network of authors and we should use that information to inform our topic model.



Term Frequencies

The fox jumped over the other fox

~~The dog ran at the foxes~~

~~The fox ran away from the dog~~

~~The cat napped on my lap~~

~~The lion napped on the hill~~

~~The cat jumped on the fox~~

FOX is considered important for DOCUMENT 1 because it is the most frequently used word in that document

Inverse Document Frequencies

The fox jumped over the other fox

The dog ran at the foxes

The fox ran away from the dog

The cat napped on my lap

The lion napped on the hill

The cat jumped on the fox

THE is considered unimportant because it can be found in all documents

Inverse Document Frequencies

The fox jumped over the other fox

The dog ran at the foxes

The fox ran away from the dog

The cat napped on my lap

The **lion** napped on the hill

The cat jumped on the fox

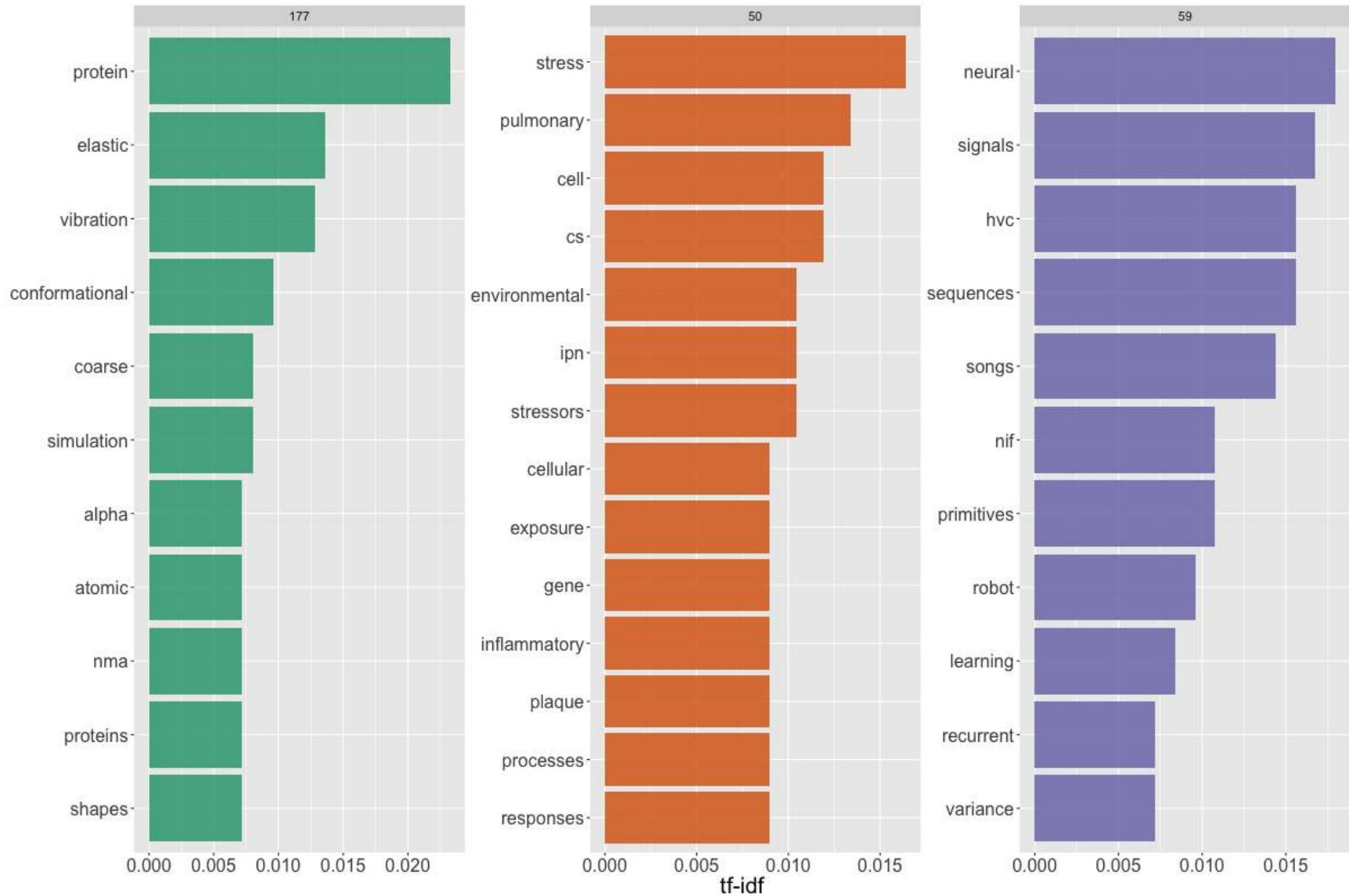
LION is considered important because it is only found in one document

Which document is most similar to document 1?

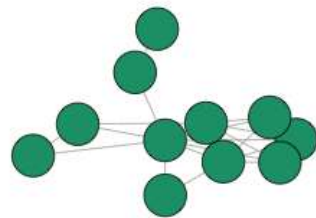
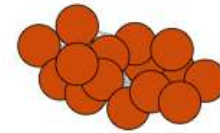
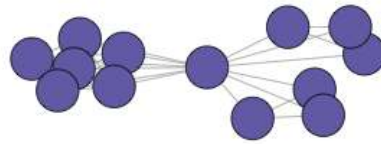
	cat	dog	fox	foxes	hill	jumped	lap	lion	napped	ran
doc_1	0	0	2	0	0	1	0	0	0	0
doc_2	0	1	0	1	0	0	0	0	0	1
doc_3	0	1	0	1	0	0	0	0	0	1
doc_4	1	0	0	0	0	0	1	0	1	0
doc_5	0	0	0	0	1	0	0	1	1	0
doc_6	1	0	1	0	0	1	0	0	0	0

TF-IDF Categories: 3 Largest Components

Let's run a TF-IDF on the three largest components and see what words differentiate them from one another

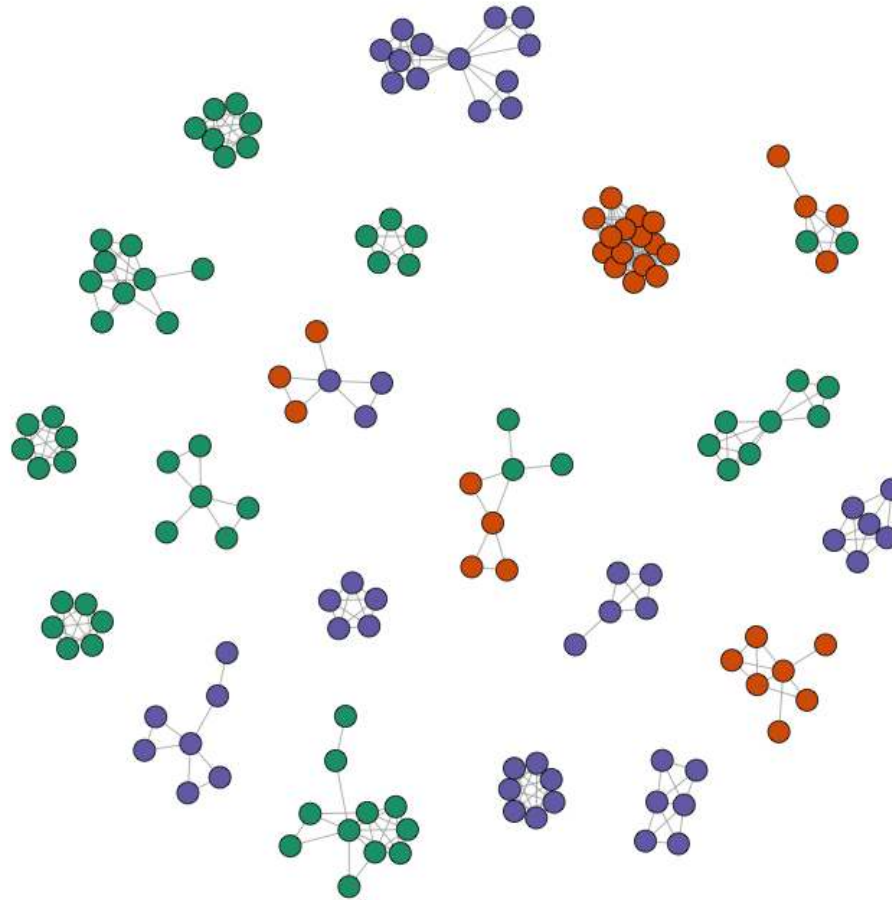


3 Largest Components: TF-IDF

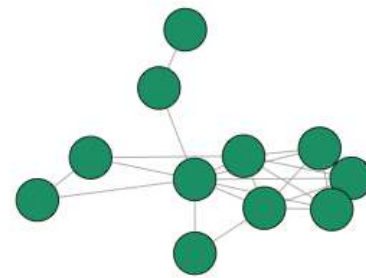
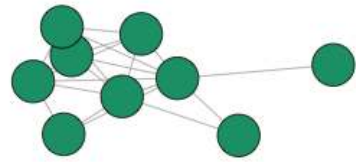


20 Largest Components: TF-IDF

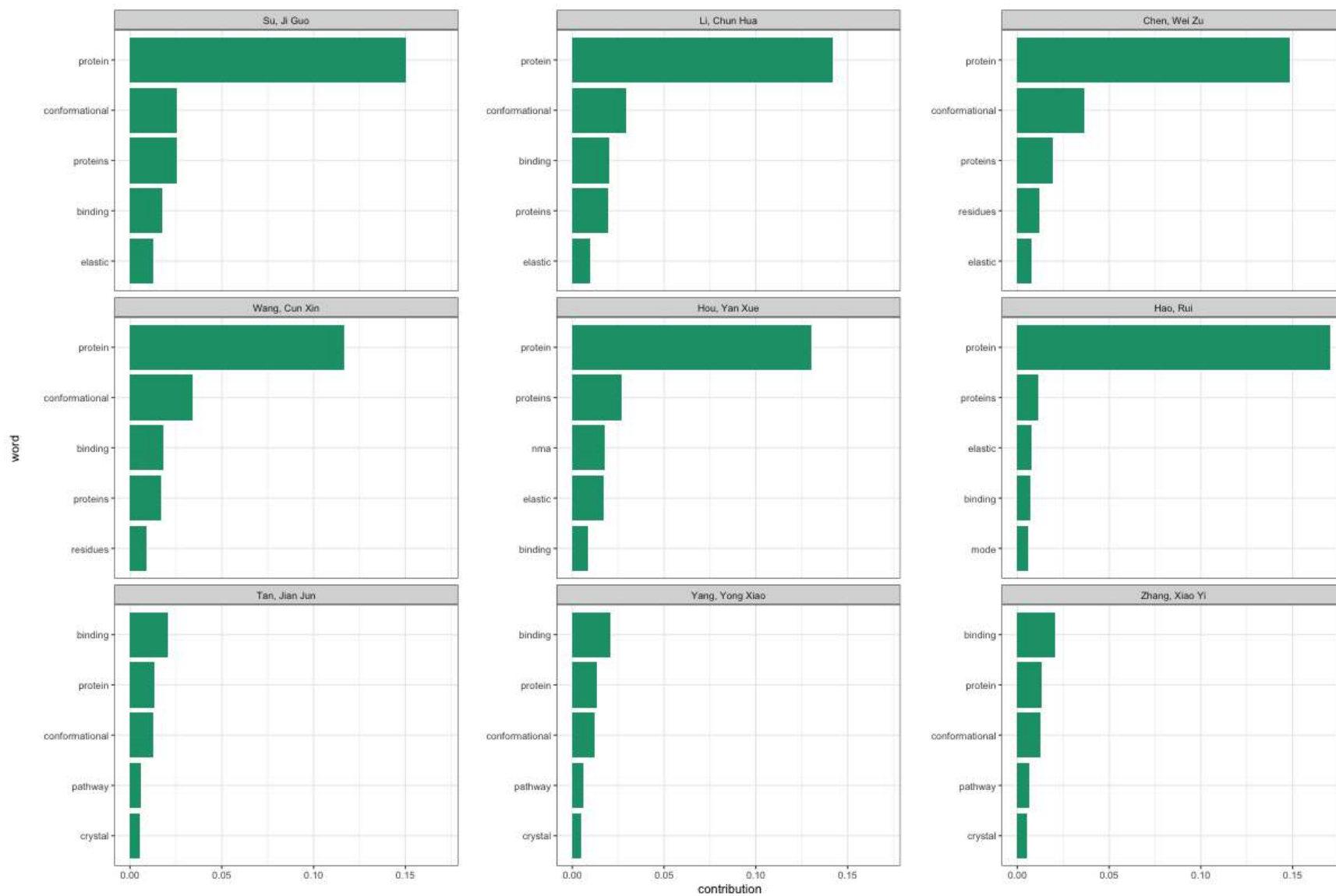
Let's use the words associated with these three largest components and compare them to the words of all the other components. This will tell us which groups are similar to the three groups we ran the initial TF-IDF on. This comparison can be done with cosine similarity because the underlying data structure of the text data is a matrix.



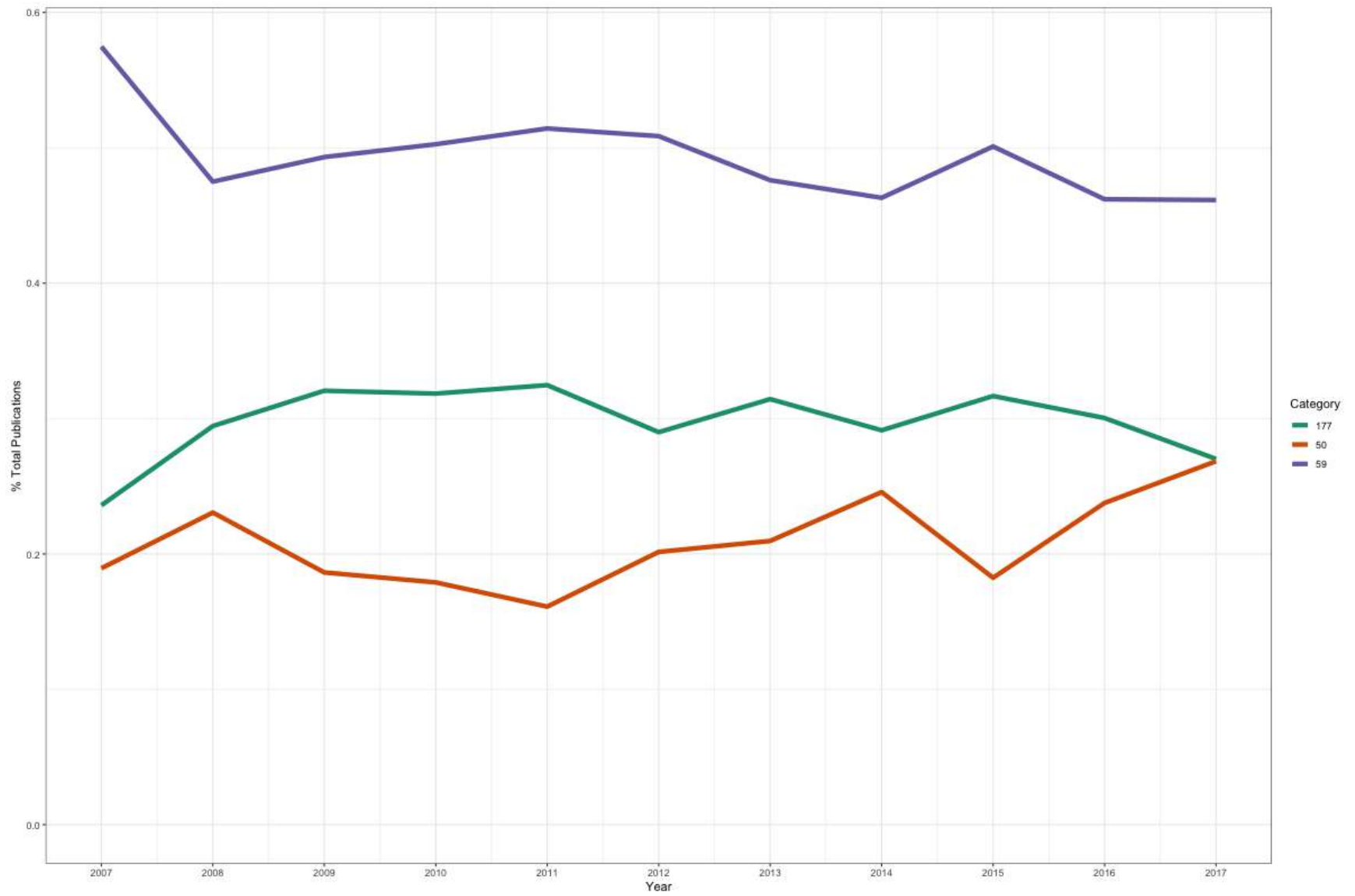
Two Similar Components

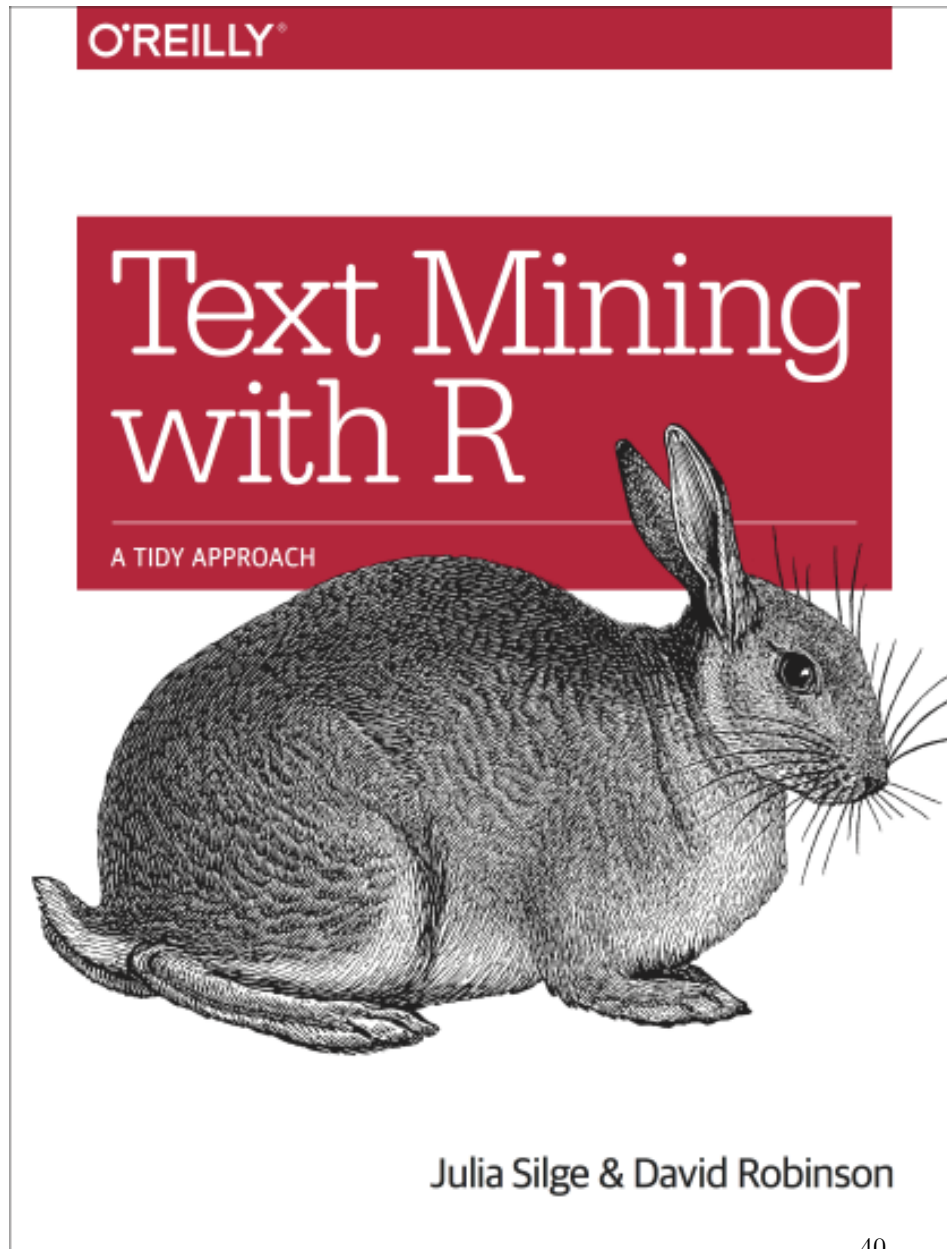


Why are these two components similar?



How have these categories changed overtime?







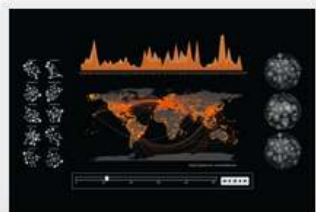
Network Science Tutorials

This page contains information about the most recent versions of several network tutorials that I have developed and frequently update. The tutorials come from workshops and invited talks I give for students, colleagues, and computationally curious bystanders. You can also find some of these materials (and other interesting bits and pieces) on my [GitHub page](#).

If you find the materials on this site to be useful, please cite them in your work. This helps me (and the computational research community) to make the case that the open publishing of digital materials, data, and code is a meaningful academic contribution.

If you want to invite me to give a talk or a workshop at your institution, email workshop@ognyanova.net.

Static and dynamic network visualization with R



This is a comprehensive tutorial on network visualization with R. It covers data input and network formats, parameters and layouts for one-mode and bipartite graphs; interactive and animated visualizations, temporal networks and visualizing networks on geographic maps.

Most recent version: 06/2018 (Polnet Conference)
Downloads: [Web version](#) | [Code & data](#) | [PDF tutorial](#).
Translations: [French](#) (L. Beauvoit), [Russian](#) (see article)

Search ... 

Katherine Ognyanova




Asst. Prof. Rutgers SC&I
Postdoc at the Lazer Lab:
Northeastern & Harvard.
PhD in Communication:
USC Annenberg School.



E-mail: kateto@ognyanova.net

Recent Tweets

 @Ognyanova and @baileyfosdick will be offering workshops in visualization and latent space models, respectively.

dc2018.satrdays.org

