Computational Text Analysis

Summer Institute in Computational Social Science

Oxford University, June 2019

Taylor W. Brown

TOPIC MODELS

What is a Topic Model?

An automated procedure for coding the content of texts (including very large corpora) into a set of meaningful categories, or "topics."

A generative model that allows sets of observations (texts) to be explained by unobserved groups (topics) that explain why some parts (words) of the data are similar.

How do Topic Models work?

Document: a bag-of-words produced according to a mixture of themes or topics that the author of the text intended to discuss.

Topic: a distribution over all observed words in the corpus.

Words strongly associated with the document's dominant topics have a higher chance of being selected and placed in the document bag (i.e. a higher chance of appearing in the document).

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Latent Dirichlet Allocation (LDA)

Objective: given a corpus of texts and a specified number of topics, find the parameters that likely generated it.

Primary input: text, and the number of topics to uncover.

Simplified process: Repeatedly pick a topic, then a word within that topic, and place them in the bag-of-words representing the document until a document is complete (i.e. "inference" or the reverse-engineering of an author's original intent).

Output: word distributions for each topic, topic distributions for the corpus.

What Topic Models are not…

"Content analysis should begin where traditional modes of research end. The [person] who wishes to use content analysis for a study of the propaganda of some political party, for example, should steep [themselves] in that propaganda. Before he begins to count, he should read it to detect characteristic mechanisms and devices. He should study the vocabulary and format. He should know the party organization and personnel. From this knowledge he should organize his hypotheses and predictions. At this point, in a conventional study, he would start writing. At this point, in a content analysis, he is, instead, ready to set up his categories, to pretest them, and then to start counting."

Harold Lasswell et al. (1952, p. 65)

What Topic Models are not…

"Seen in this light, it is useful to think about topic models not as providing an automatic text analysis program but rather as providing a lens that allows researchers working on a problem to view a relevant textual corpus in a different light and at a different scale."

Mohr & Bogdanov (2013)

Structural Topic Models (STM)

"A general framework for topic modeling with document-level covariate information. The covariates can improve inference and qualitative interpretability and are allowed to affect topical prevalence, topical content or both."

http://www.structuraltopicmodel.com/

R package: STM

WORD EMBEDDING MODELS

What is word embedding?

A method of text analysis that results in a matrix of word vectors, in which words used in similar contexts are closer together in vector space than words used in different contexts.

https://www.tensorflow.org/tutorials/representation/word2vec

https://projector.tensorflow.org/
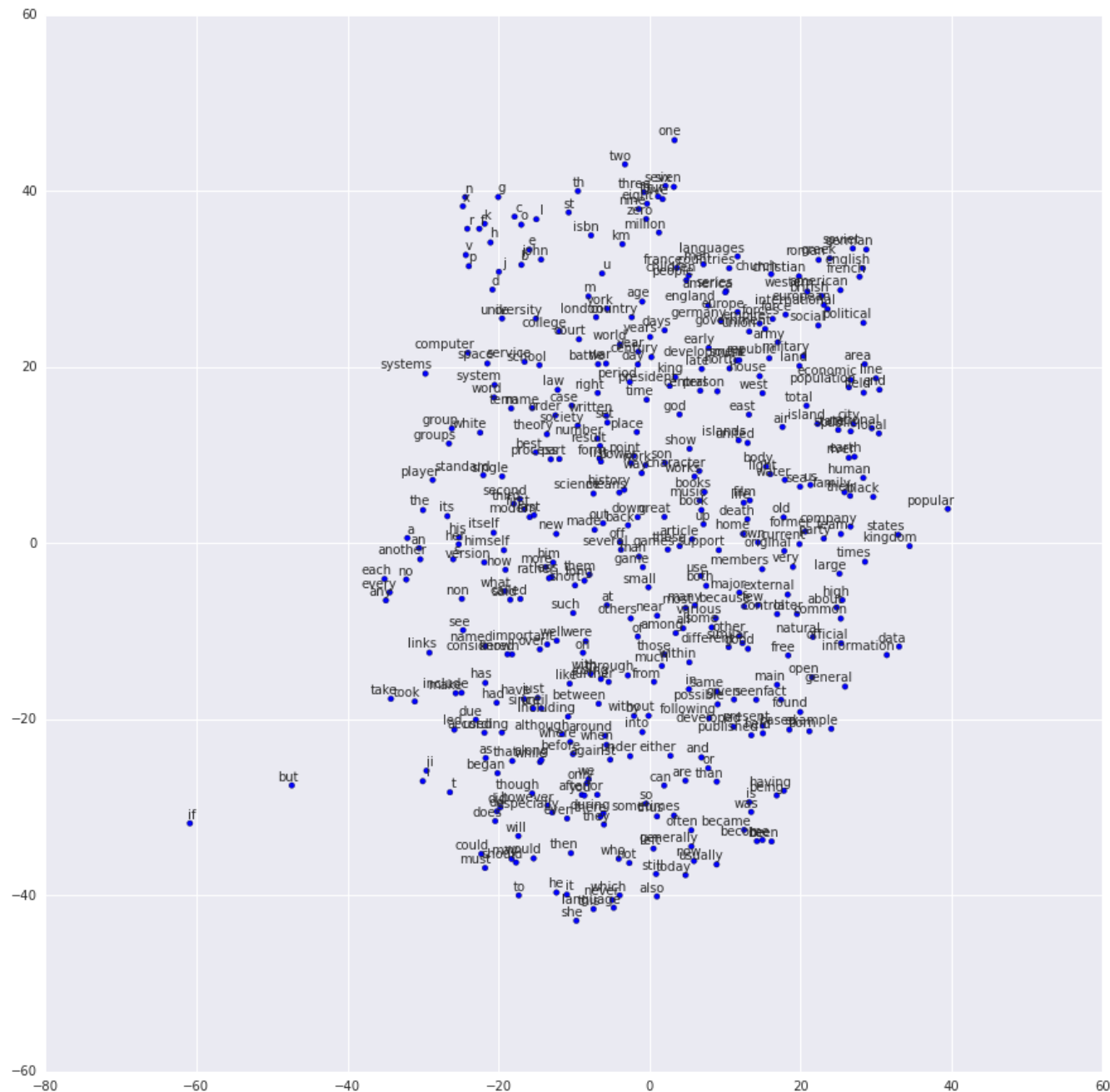
How does it work?

Two common approaches (there are others)

continuous bag-of-words (CBOW)

predict a word given its context, where context is
defined by the research a n number of words
surrounding the target word (usually ~8).

skip-gram

predict a context given a word

Q: What is the Word2Vec method for mapping semantic space?

A: skip-gram with negative sampling

How does it work?

First, take a word/concept in your training corpus "target" and a number of words that lie close to it "context."

For example, take the following passage from Catch-22:

> *"People who had hardly noticed his resemblance to Henry Fonda before now never ceased discussing it, and there were even those who hinted sinisterly that Major Major had been elevated to squadron commander because he resembled Henry Fonda. Captain Black, who had aspired to the position himself, maintained that Major Major really was Henry Fonda but was too chickensh*t to admit it."*

Let's say that the character name 'Major Major' is our target, and the five words on either side of it are its context... so:

> *"People who had hardly noticed his resemblance to Henry Fonda before now never ceased discussing it, and there were even those who hinted sinisterly that Major Major had been elevated to squadron commander because he resembled Henry Fonda. Captain Black, who had aspired to the position himself, maintained that Major Major really was Henry Fonda but was too chickenshit to admit it."*

How does it work?

To map the semantic space of a text corpus, w2v uses a technique called 'skip-gram with negative sampling.'

(1) take a word/concept in your training corpus "target" and a number of words that lie close to it "context."

(2) represent each of these words by a vector (a list of numbers); to begin with, these vectors can be random.

The aim here is to get the vector of our target and the vectors of its context to be close to one another in vector space (approximated by taking the dot product of the vectors), so:

(3) take the target and context vectors and pull them together by a small amount to make them closer.

This is accomplished by maximizing the predicted probability of the two words co-occurring (given by a logit transformation of the dot product of the target and the context word).

Additionally, we want our target word to be further away from words it is rarely used in context with (i.e. we want their vectors to be dissimilar). To accomplish this we:

(4) randomly sample words from the rest of the corpus (i.e. words outside of our target's context).

(5) push the vectors of these random words a little further away from the vector of our target.

Here we minimize the logit of the dot product of target and each non-context sample word.

How does it work?

Better with large and/or topically consistent, and non ambiguous corpora.

Why does it matter?

Given that the semantic description of words in a corpora are represented as numeric vectors in w2v, once all words have been mapped into the vector space, it becomes possible to use vector math to find words that have similar semantics or more complex relationship.

For example, an often cited instance of vector math depicting conceptual relationships with w2v is:

[the vector for 'king'] - [the vector for 'man'] + [the vector for 'woman'] = [the vector for 'queen'].

How does it differ from topic modeling?

As Ben Schmidt puts it

"A topic model aims to reduce words down some core meaning so you can see what each individual document in a library is really about. Effectively, this is about getting rid of words so we can understand documents more clearly.

WEMs do nearly the opposite: they try to ignore information about individual documents so that you can better understand the relationships between words."

True, but topics also have distributions across the corpus, so there's more to it than that…

How does it differ from topic modeling?

Topic models sort words into a predetermined n of topics and do not (aim to) capture continuous relationships between words.

They don't do well at representing associations between words, or how words mediate and moderate meaning for one another, which is a strength of word embedding models, as we will see in the tutorial.

Now we'll move over to Python and go through the Wort2Vec tutorial.

You can find it in the file:
SICSSOxford2019_WordEmbedding_Tutorial.html

We'll be using the **soc_abstracts.txt** data.

NETWORK ANALYSIS

Network analysis:
a set of *relational* methods for systematically understanding connections between classes of entities.

"Structured social relationships are a more powerful source of sociological explanation than personal attributes of system members."

- Harrison White

"The way that you use words in relation to one another is a powerful source of sociological explanation than the individual meaning of words."

Networks as variables

1. Are people who hang out with Republicans more likely to be Republicans?
2. Are negative words more likely to be associated with authority words?
3. Do central words signify document meaning?

Networks as structures

1. What word network patterns generate the spread of ideas most quickly?
2. How do ideas evolve out of consistent relational word usage?

Connectivist: networks matter because of what flows through them

1. Diseases (STD, Ebola, etc)
2. Information (fake news, job opportunities, etc)
3. Behaviors (self-harming behavior, happiness, anxiety, rap, etc)

Positionalist: networks matter because of the way they capture role behavior

1. What word network patterns generate the spread of ideas most quickly?
2. How do ideas evolve out of consistent relational word usage?

Node

the entities being connected

Attributes (e.g. gender, class, polarity, word type, etc)
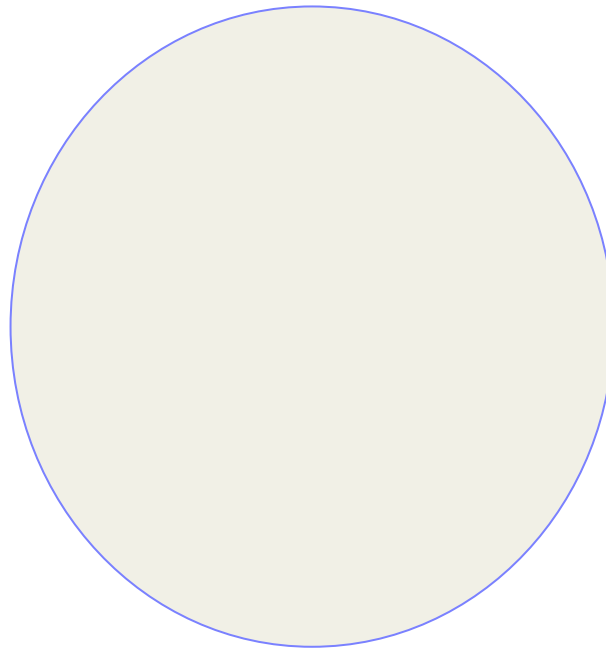
Taylor

Chris

Aidan

Friedo

Marcus

Edge

a type of relationship between nodes

Attributes (e.g. binary, valued, directed, undirected, time)
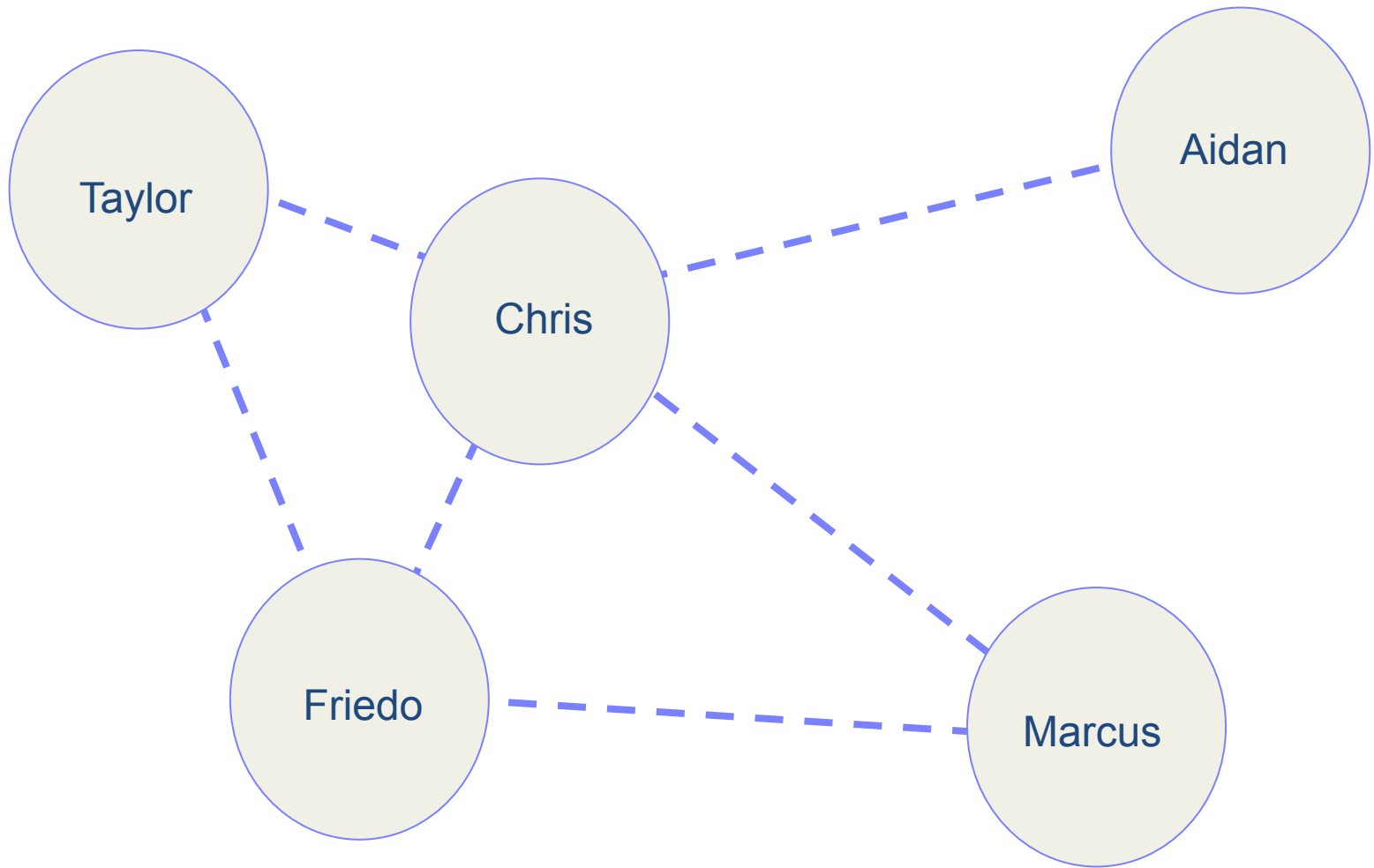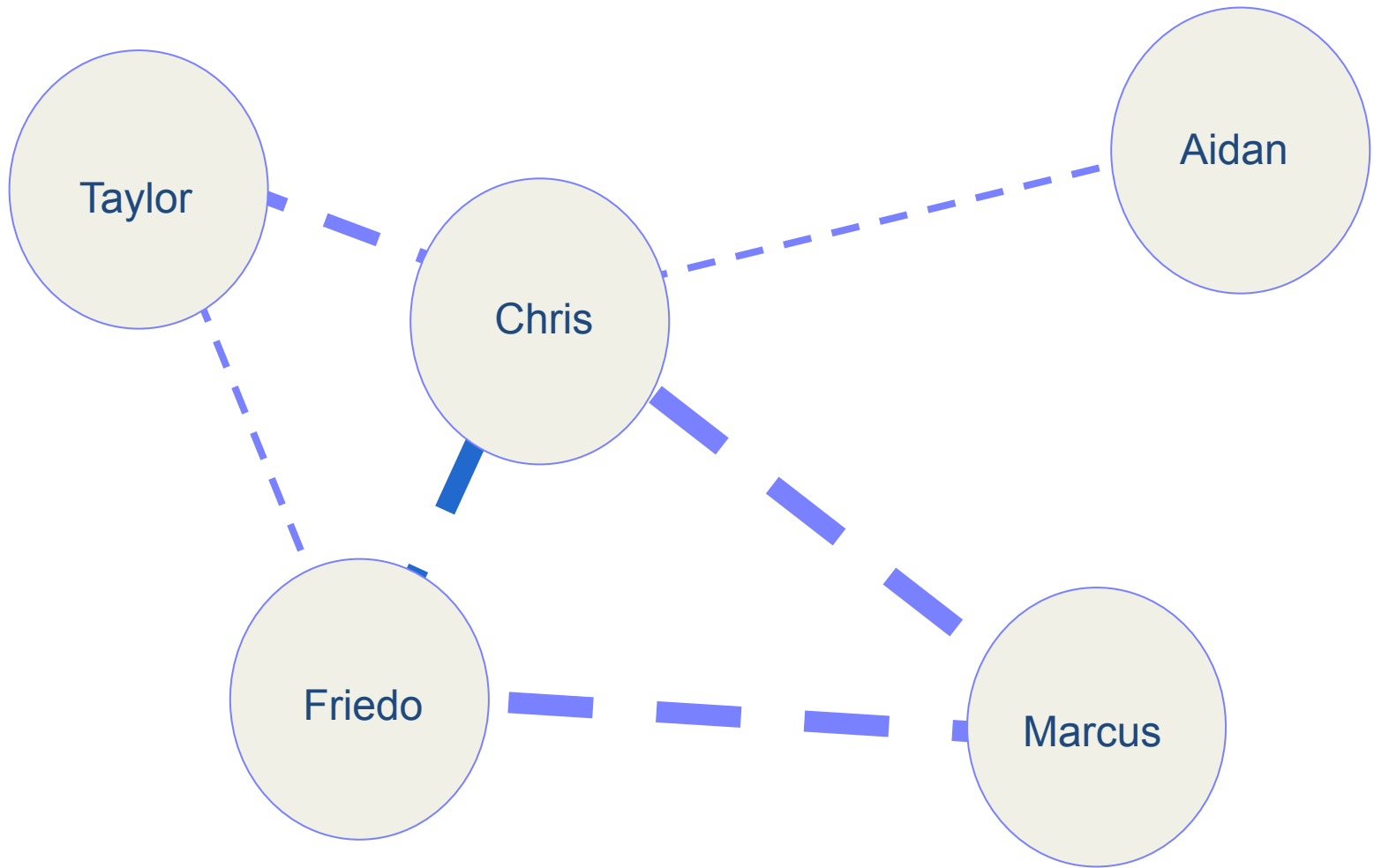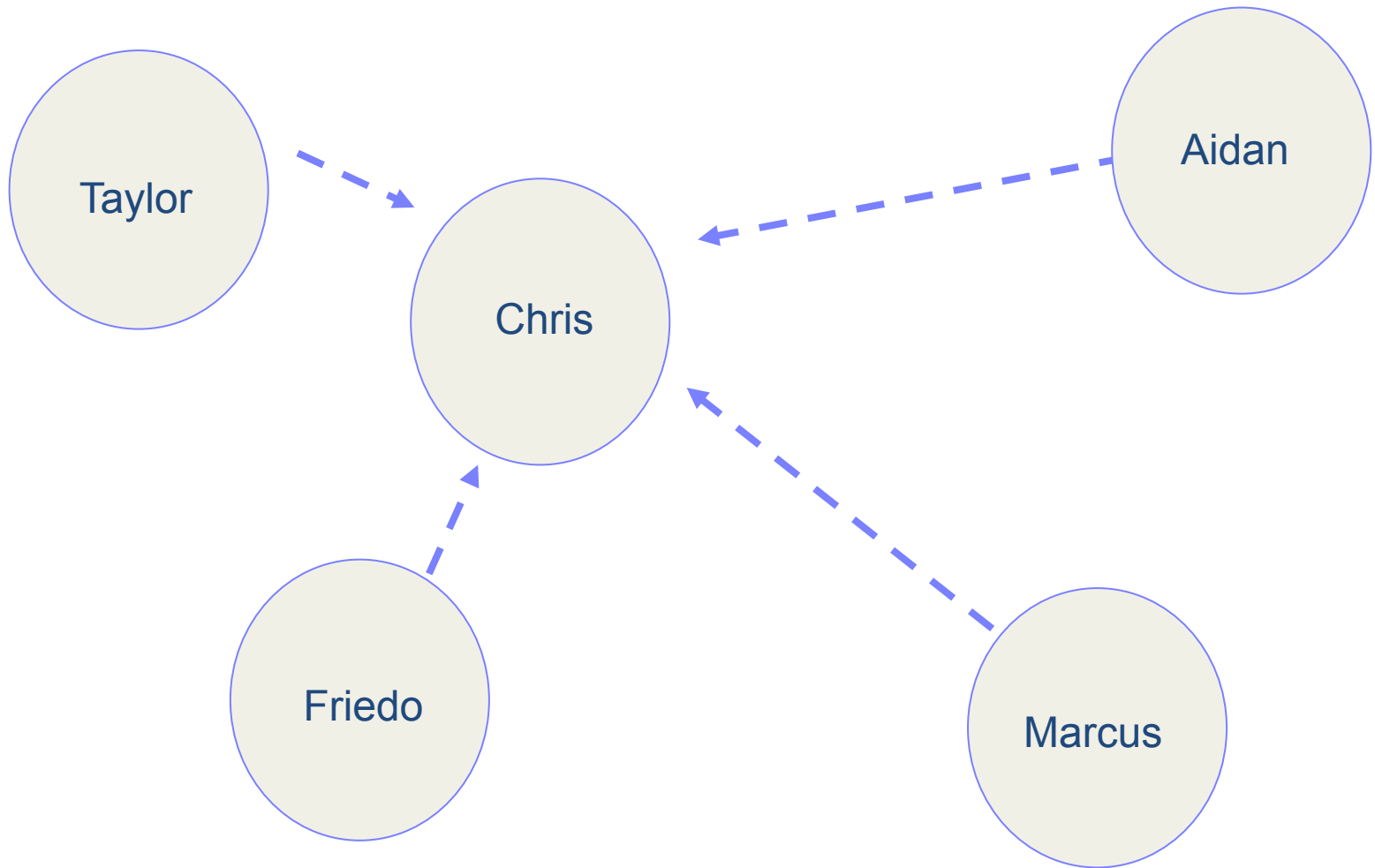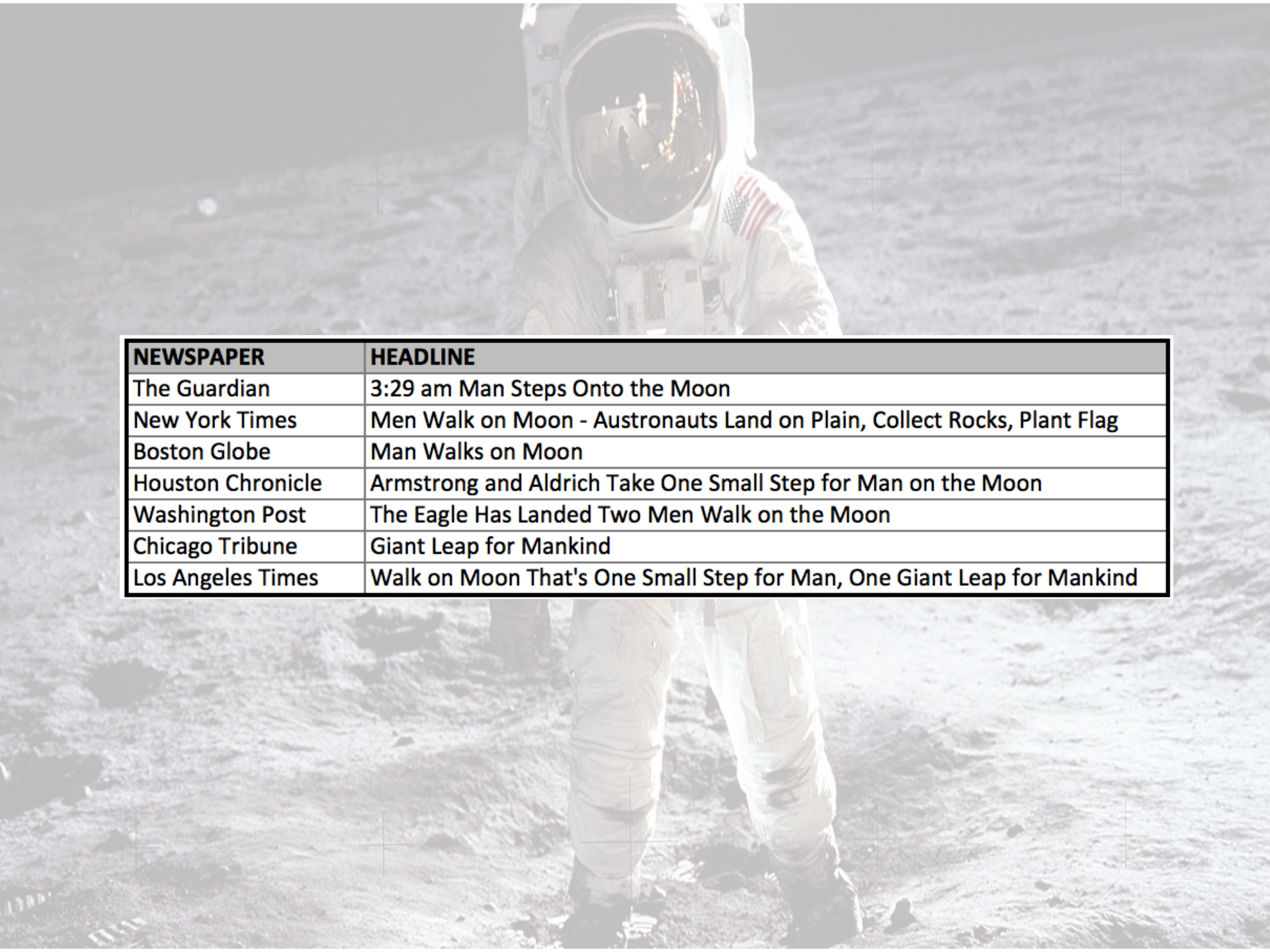
Graph theory notation:
G(V,E)

NEWSPAPER	HEADLINE
The Guardian	3:29 am Man Steps Onto the Moon
New York Times	Men Walk on Moon - Austronauts Land on Plain, Collect Rocks, Plant Flag
Boston Globe	Man Walks on Moon
Houston Chronicle	Armstrong and Aldrich Take One Small Step for Man on the Moon
Washington Post	The Eagle Has Landed Two Men Walk on the Moon
Chicago Tribune	Giant Leap for Mankind
Los Angeles Times	Walk on Moon That's One Small Step for Man, One Giant Leap for Mankind
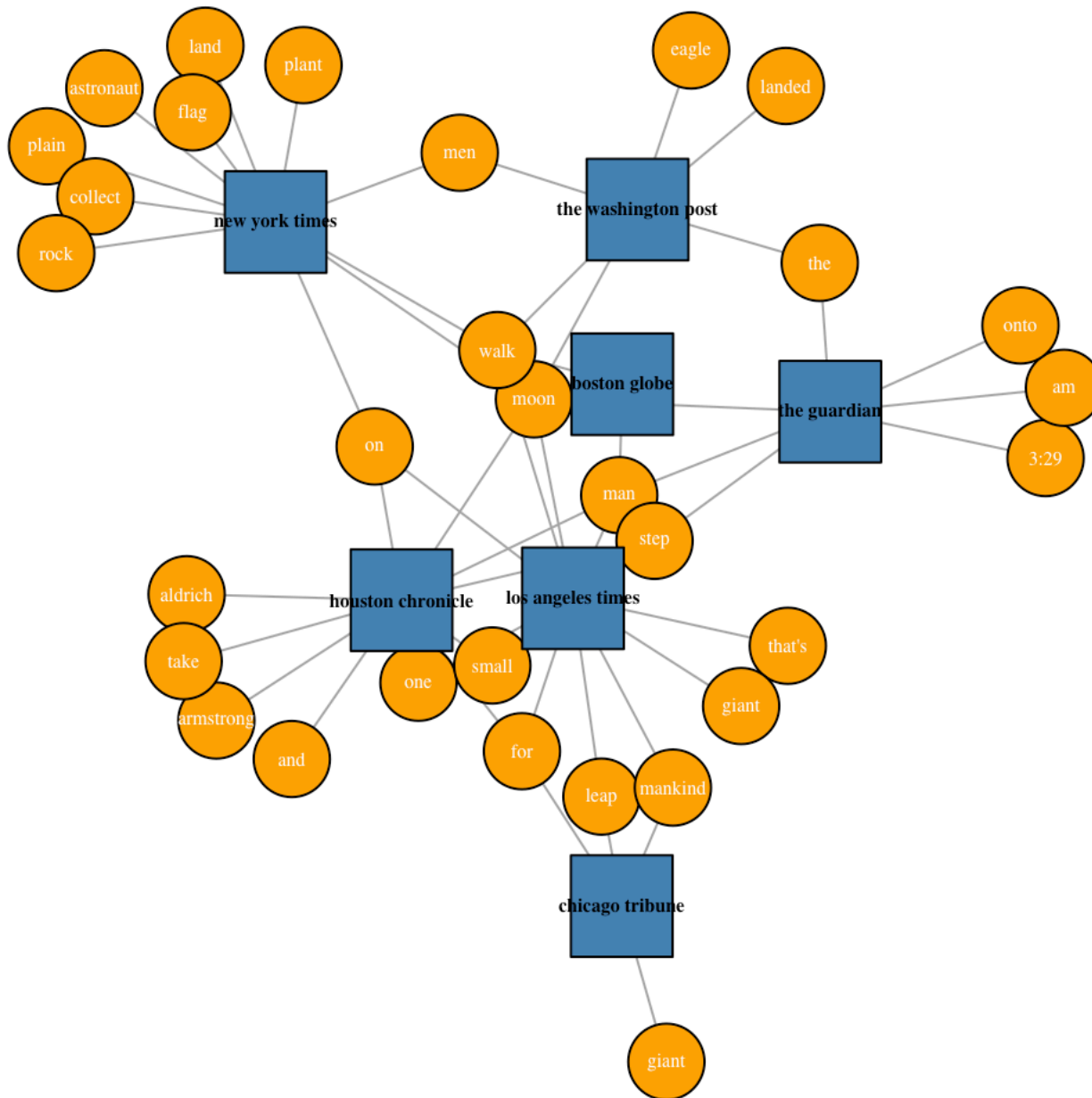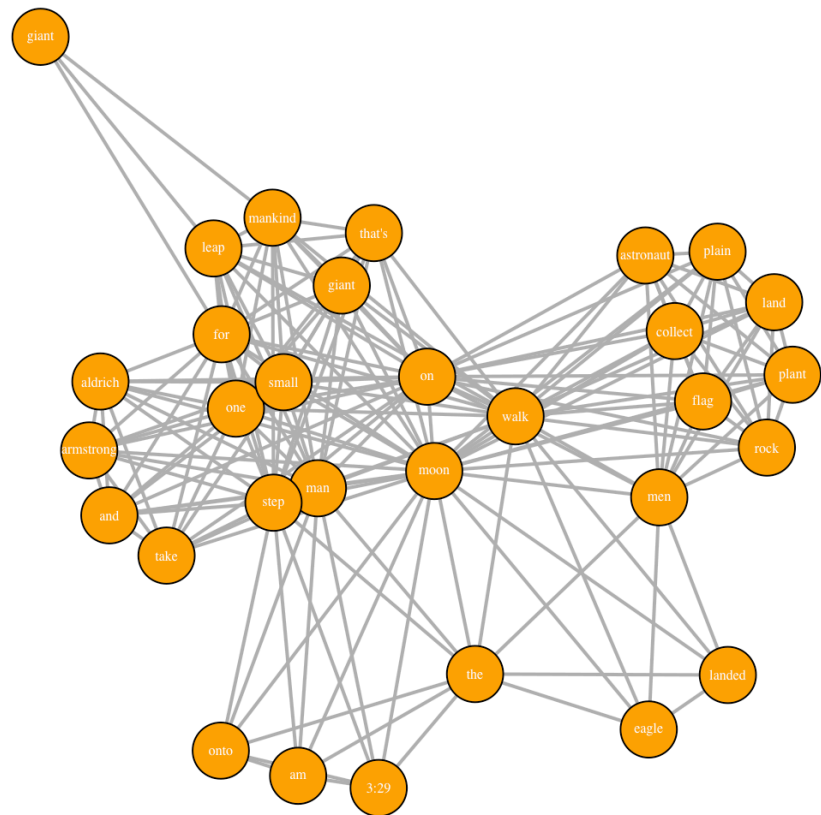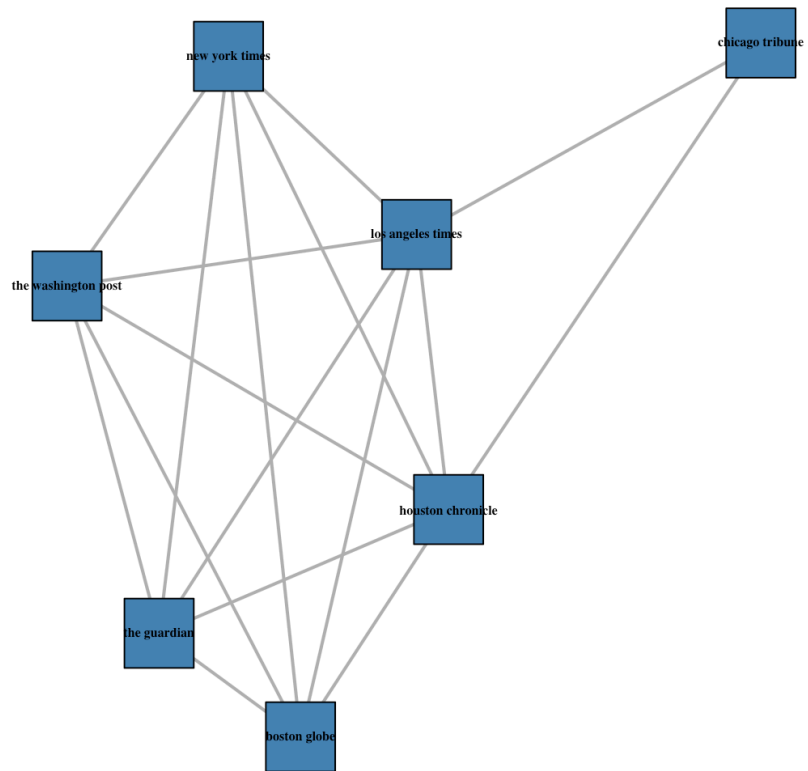
Centrality

1. Closeness: sum of geodesic distances from a word to all others
2. Eigen: number of ties that a word shares with other word, but weighted by the centrality of the alters
3. Betweenness: how often a word falls along the shortest path between two other words

Cohesion

1. Disperse (words connected at relatively equal rates), Clustered (certain words more likely to occur together), Core-periphery (with a large tight group, and a few words that are disconnected)
2. Density: number of ties in the net- work relative to the number of ties possible
3. Cluster: cohesive subgroups where in words experience stronger or more frequent pairwise relationships
   a. Transitivity: property of a network that describes a particular pattern of triples of actors--if A-B, and B-C, then A-C
      i. measured by clustering coefficient: the average fraction of pairs of alters of a actor which are also alters of each other.
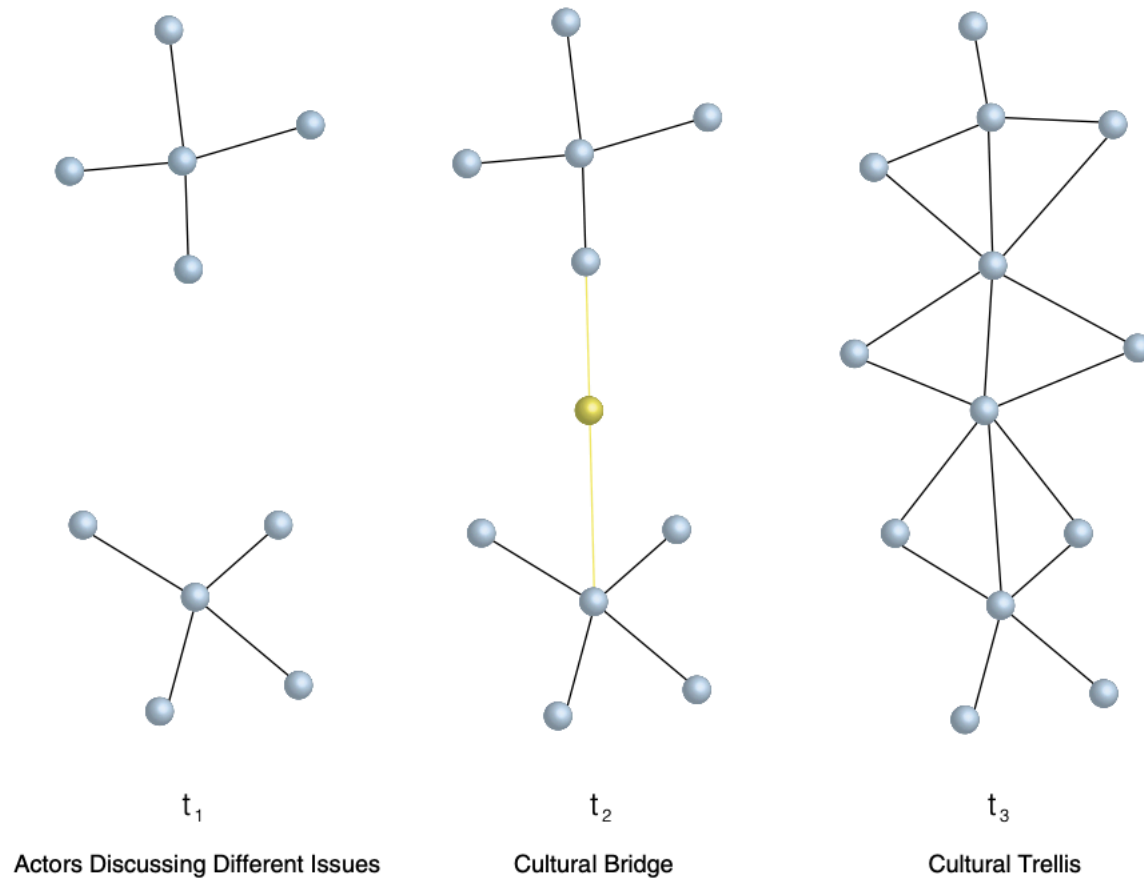
Fig 1. Hypothetical cultural network in which nodes represent actors engaged in conversation about an advocacy issue and edges between them describe similarities in
the content of their messages.

Bail (2016). Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. PNAS, 113 (42), pp. 11823–11828.

"This study reveals that the entry into World War I (WWI) in 1917 indexed the decisive transition to the modern period in American political consciousness, ushering in new objects of political discourse, a more rapid pace of change of those objects, and a fundamental reframing of the main tasks of governance."
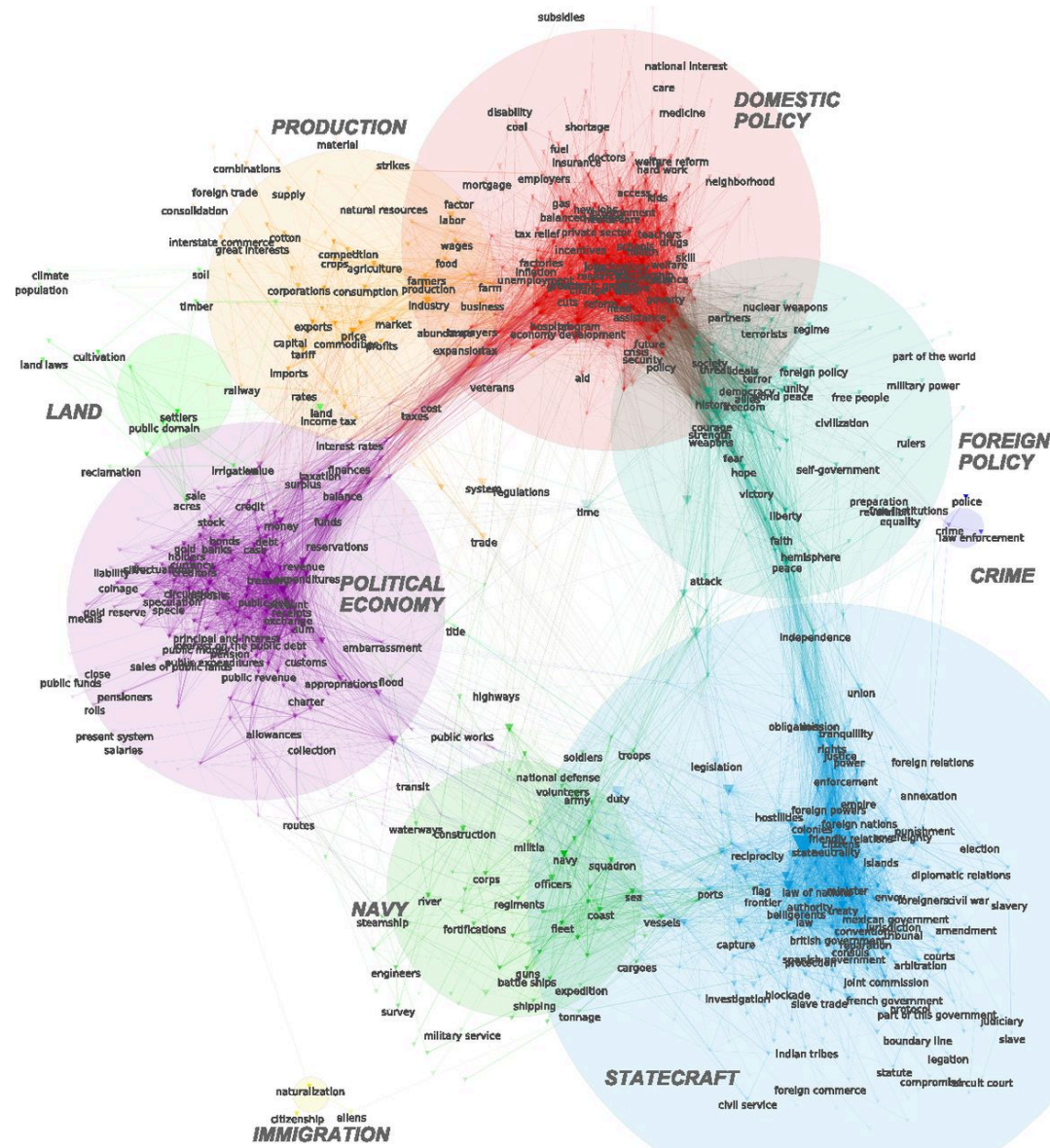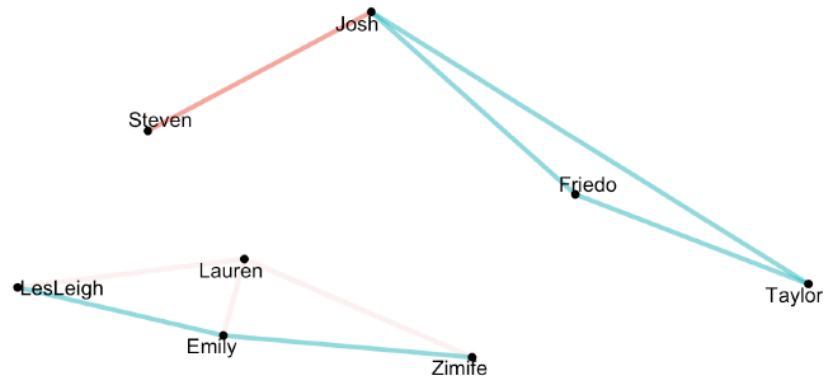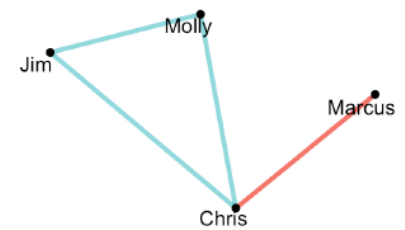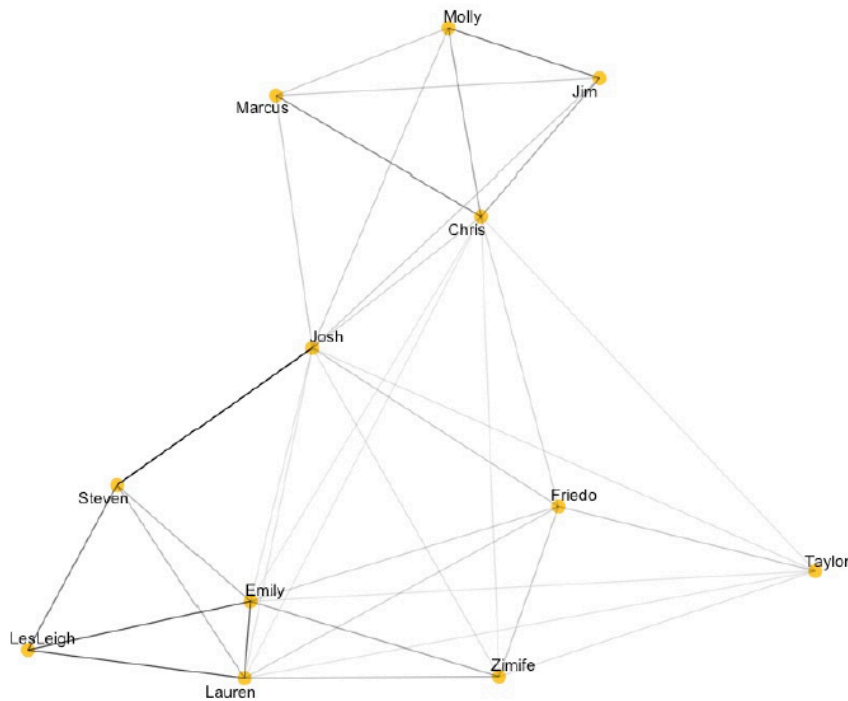
Fig 3. The global network structure of the SoU, 1790–2014. The Louvain community detection.

Rule, Cointet, Bearman (2015). *Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014.* PNAS, 112 (35), pp. 10837-10844

`textnets` package

newly developed R package that
combines automated text analysis with
methods from network analysis.

Speaker	Sentence	Sentiment
Chris	The strong women and hypocritical Disney are excellent.	positive
Marcus	The hypocritical women are obnoxious.	negative
Friedo	Strong men are everywhere.	positive
Taylor	Strong men of morals are needed in the government.	positive
Jim	The hypocritical Disney is a travesty.	negative
Molly	I think hypocritical Disney should be quiet.	negative
Josh	Where are all of the hypocritical actors and strong men?	negative
Steven	Just ordinary actors doing their thing.	neutral
Zimife	Get ready for a strong storm.	positive
Lauren	This is just an ordinary storm and an strong workday.	neutral
Emily	An ordinary workday and a strong storm.	neutral
LesLeigh	This is an ordinary workday.	neutral

Now, if we have time, we'll move over to R for the tutorial of the textnets package

See the folder called **textnetsMaterial**

GREP cheatsheet: http://www.rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf

Tidy Text Mining with R: https://www.tidytextmining.com/

`stringr` cheatsheet: http://edrub.in/CheatSheets/cheatSheetStringr.pdf

`quanteda` package: https://cran.r-project.org/web/packages/quanteda/quanteda.pdf

`stringi` package: https://cran.r-project.org/web/packages/stringi/stringi.pdf

`stringr` vignette: https://cran.r-project.org/web/packages/stringr/vignettes/stringr.html

`stringi` vs `stringi`:
https://www.r-bloggers.com/strung-out-on-string-ops-a-brief-comparison-of-stringi-and-stringr/

Gensim (Python) https://radimrehurek.com/gensim/index.html

wordVectors: https://github.com/bmschmidt/wordVectors