Topic Modeling:

from a Network Perspective

Cindy Cook

Network Analysis: Presentation

26 April 2016

Outline

1. Overview of Topic Modeling
2. Replication of Paper
3. Data Sets
4. Extension

Topic Modeling: The General Idea

1. Obtain documents, D and have an idea as to the number of topics K within D.
2. Stem and remove stop words.
3. Create the term-frequency matrix.
4. Apply model: LDA, pLSA, etc...
5. Obtain:
   a) For each d, a topic distribution.
   b) For each topic k, a word distribution.

Example:

$d1$:I like to eat fish for dinner.

$d2$:I only eat ice cream after dinner.

$d3$:I own a pet fish named Nemo.

Term-Frequency Matrix / Topic/Word Distributions

	$d1$	$d2$	$d3$	$k1$	$k2$	$k1$	$k2$
cream	0	1	0	0.0833	0	0.0825	0.0714
dinner	1	1	0	0.1666	0	0.1480	0.1597
eat	1	1	0	0.1666	0	0.2710	0.0369
fish	1	0	1	0.0833	1	0.1509	0.1568
ice	0	1	0	0.0833	0	0.0935	0.0604
like	1	0	0	0.0833	0	0.0387	0.1150
name	0	0	1	0.0833	0	0.0255	0.1283
nemo	0	0	1	0.0833	0	0.0809	0.0729
own	0	0	1	0.0833	0	0.0253	0.1285
pet	0	0	1	0.0833	0	0.0837	0.0701
$k1$	0.5010	0.5051	0.4939				
$k2$	0.4990	0.4949	0.5061				
$k1$	0	0	0.8				
$k2$	1	1	0.2				

A Network Prespective: TopicMapping

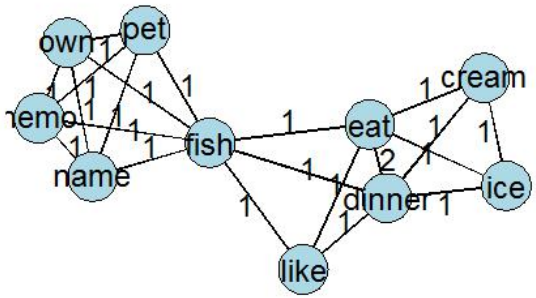
*Before step 4:

Bipartite Graph

Term-Frequency Matrix

	d1	d2	d3
cream	0	1	0
dinner	1	1	0
eat	1	1	0
fish	1	0	1
ice	0	1	0
like	1	0	0
name	0	0	1
nemo	0	0	1
own	0	0	1
pet	0	0	1

Projected



Estimates

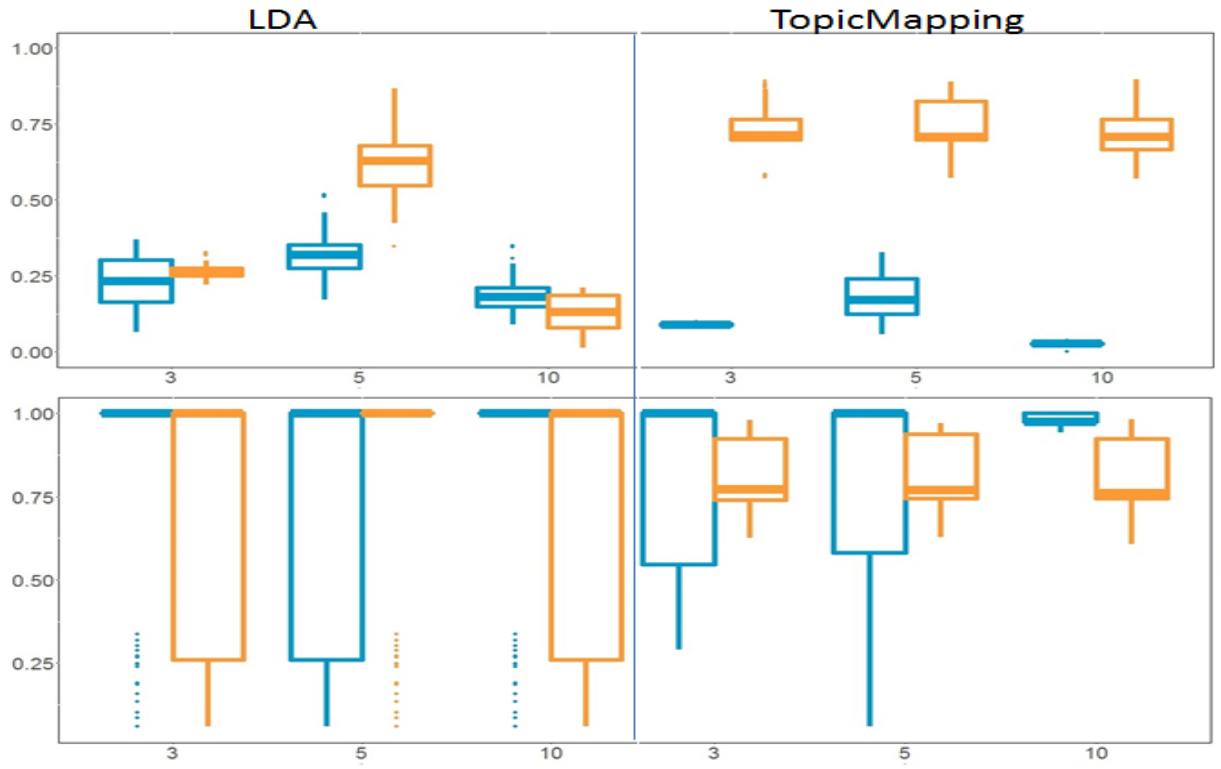
Topic/Word Dists.

	k1	k2
cream	0.1667	0
dinner	0.1667	0
eat	0.1667	0
fish	0.1667	1
ice	0.1667	0
like	0.1667	0
name	0	0.25
nemo	0	0.25
own	0	0.25
pet	0	0.25

*Replication Study:

Data

1. Web of Science: Each document includes the title and abstract of an article from different fields including economics, astronomy, psychology, biology, and math.
2. Indeed.com: Each document includes past work experience for an individual in teaching, graphic, architecture, accounting, and nursing.

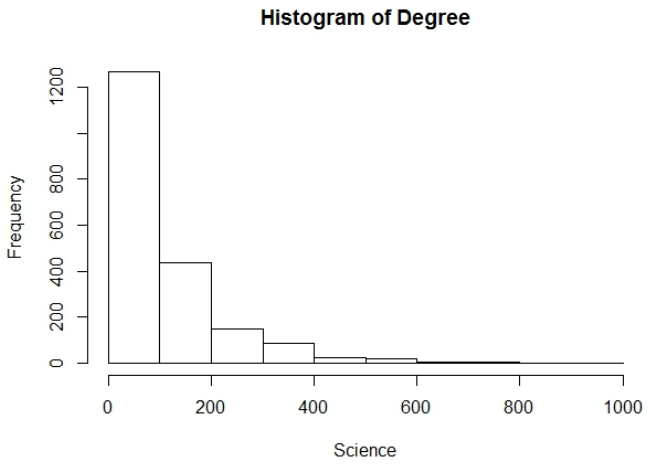


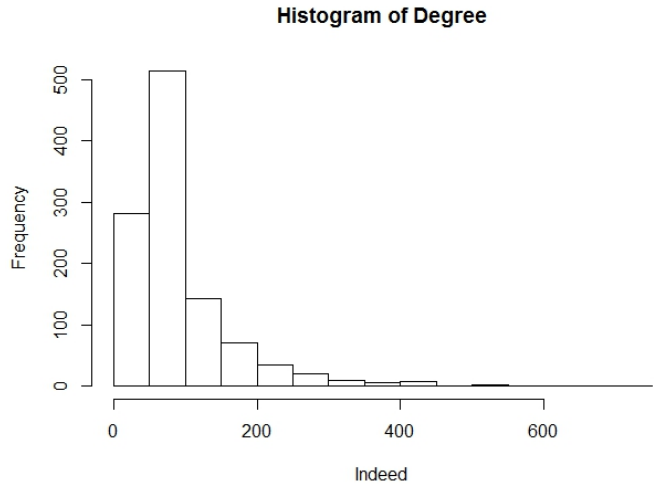
Network Comparison: Descriptive

Modularity:

Clustering Type	Science	Indeed
InfoMap	0.3884	0.3881
Fast-Greedy	0.3398	0.3358
Eigenvalue	0.3624	0.3616
Louvain	0.3936	0.3904
Walktrap	0.3257	0.3431

Density, Transitivity, Betweenness Centrality, Degree Assortativity, Degree Distributions:





Network Comparison: Extension

*ERGM

	Science				Indeed			
	Estimate	Std. Error	MCMC%	p-value	Estimate	Std. Error	MCMC%	p-value
edges	-3.3591	0.0040	0	< 0.0001	-2.9925	0.0063	0	< 0.0001
isolate	-Inf	0	0	< 0.0001	-Inf	0	0	< 0.0001
homophily	3.3696	0.0077	1	< 0.0001	3.3616	0.0122	0	< 0.0001

*QAP

	Science				Indeed			
	Estimate	$\Pr(\leq b)$	$\Pr(\geq b)$	$\Pr(\geq \lvert b \rvert)$	Estimate	$\Pr(\leq b)$	$\Pr(\geq b)$	$\Pr(\geq \lvert b \rvert)$
intercept	0.0463	1	0	0	0.0621	1	0	0
Homophily	0.9537	1	0	0	0.9379	1	0	0
Adj. R^2	0.2406				0.245			

*CUG

-Conditional on each network's dyad distribution.

Modularity

Transitivity