



Graph Learning and Misinformation

Captain Iain Cruickshank
Dr. Kathleen M. Carley
12 June 2019



Carnegie Mellon

Center for Computational Analysis of
Social and Organizational Systems
<http://www.casos.cs.cmu.edu/>



Misinformation is a Growing Problem

- Fake news is now viewed as one of the greatest threats to **democracy, justice, public trust, freedom of expression, journalism** and **economy (Shu et al 2019)**.
- Credited with significant measurable impacts
 - Influencing elections (US elections 2016, and 2018), UK Brexit vote, and recent Indian Elections
 - Has also caused significant economic impacts (e.g. false tweet of Barrack Obama being injured in explosion cause a drop \$130 Billion in Stocks)




June 2019

CASOS Center

2






Carnegie Mellon
Institute for
SOFTWARE
RESEARCH

Misinformation is also hard to characterize and define


	Authenticity	Intention	News?
Fake news	False	Bad	Yes
False news	False	Unknown	Yes
Satire news	Unknown	Not bad	Yes
Disinformation	False	Bad	Unknown
Misinformation	False	Unknown	Unknown
Rumor	Unknown	Unknown	Unknown

X. Zhou, R. Zafarani, K. Shu, H. Liu. *Fake News: Fundamental Theories, Detection Strategies and Challenges* (2019)



CASOS


June 2019CASOS Center3



Carnegie Mellon
Institute for
SOFTWARE
RESEARCH

Misinformation is also hard to characterize and define

- Misinformation can be characterized by many means
 - Knowledge – based (e.g. fact checking)
 - Style – based (e.g. diction, writing, style)
 - Propagation – based (e.g. how something spreads across a network, co-publication, etc.)
- There is no one best way to characterize the essential elements of misinformation



CASOS

June 2019CASOS Center4



The diagram illustrates a Fake News Detection System Architecture. It is divided into two main sections: 'Fake News Collection' and 'Fake News Detection'.

Fake News Collection: This section includes a central 'DB' (Database). It receives data from several sources: 'POLITIFACT' (via a 'Fact-Checking Crawler'), 'Twitter Ads, Search Crawler', 'News Content Crawler', and 'BuzzFeed'. The 'Twitter Ads, Search Crawler' also feeds into a 'Tweet Engagement Crawler'.

Fake News Detection: This section takes input from the 'DB' and the 'Tweet Engagement Crawler'. It leads to a 'Fake News Detection' box, which then outputs to a 'Fake News Visualization' box. The visualization shows a bar chart and a line graph on a tablet screen.

Misinformation Data is Multi-Modal

Data describing the same event or persons often comes from many sources

The diagram illustrates a three-stage process for handling multi-modal data:

- Objects of interest in some unobserved, latent space:** Represented by a box containing several blue dots.
- Different observations of those objects producing multi-modal data:** Represented by three overlapping rectangles (light gray, dark gray, and orange) with arrows pointing from the latent space box to them.
- Fusion of data into a data model that best fits the data's latent structure:** Represented by a grid of blue squares with arrows pointing from the multi-modal data rectangles to it.

CASOS

June 2019

CASOS Center

6



Carnegie Mellon

ISI

Institute for
SOFTWARE
RESEARCH

Problem Statement

We would like to have means of combining multi-modal Misinformation data into a flexible data model that correctly highlights patterns of interest, trends, etc. in an unsupervised way

AP The Associated Press

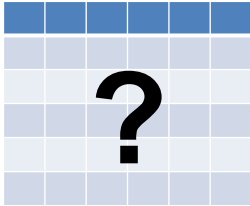
Breaking: Two Explosions in the White House and Barack Obama is injured

876 RETWEETS 32 FAVORITES

Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement

BREAKING: Obama And Hillary Now Promising Amnesty To Any Illegal That Votes Democrat

→



CASOS

June 2019

CASOS Center

7

Carnegie Mellon

ISI

Institute for
SOFTWARE
RESEARCH

Problem Statement

We would like to have means of combining multi-modal Misinformation data into a flexible data that correctly highlights patterns of interest, trends, etc.

AP The Associated Press

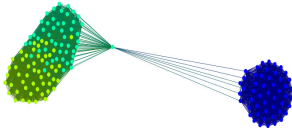
Breaking: Two Explosions in the White House and Barack Obama is injured

876 RETWEETS 32 FAVORITES

Pope Francis Shocks World, Endorses Donald Trump for President, Releases Statement

BREAKING: Obama And Hillary Now Promising Amnesty To Any Illegal That Votes Democrat

→




CASOS

June 2019

CASOS Center

8




Carnegie Mellon

ISI


Institute for
SOFTWARE
RESEARCH

Graphs Are Good Data Models

- Have emergent structures
- Allow for local heterogeneity
- Interpretable by man & machine
- Widely used in manifold learning



CASOS



June 2019

CASOS Center

9


Carnegie Mellon

ISI


Institute for
SOFTWARE
RESEARCH

Graph Learning

- The fundamental idea of graph learning is to find the best graph representation of some data
 - It could be considered as a way of approximating the manifold of the data
 - A recent survey of the field is available in Qiao et al. *Data-driven graph construction and graph learning: A review* and Brugere et al. *Network Structure Inference, A Survey: Motivations, Methods, and Applications*
- Used in everything from subspace learning, clustering, dimensionality reduction, manifold learning, metric learning, etc.



CASOS



June 2019

CASOS Center

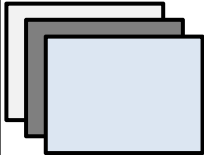
10



Carnegie Mellon


ISI
Institute for
Software
Research

Graph Learning with Multi-Modal Data



Measure different modalities of the phenomenon

CASOS



June 2019

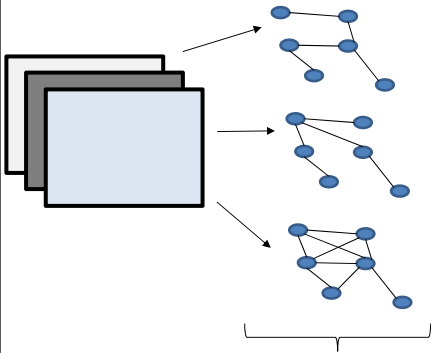
CASOS Center

11

Carnegie Mellon


ISI
Institute for
Software
Research

Graph Learning with Multi-Modal Data



Compute starting graphs of each mode of the data

CASOS



June 2019

CASOS Center

12



Carnegie Mellon
ISI Institute for Software Research

Graph Learning with Multi-Modal Data

Compute starting graphs of each mode of the data

Determine the inclusion of parts of each the graphs into a final fused manifold graph

CASOS

June 2019 CASOS Center 13

Carnegie Mellon
ISI Institute for Software Research

Graph Learning with Multi-Modal Data

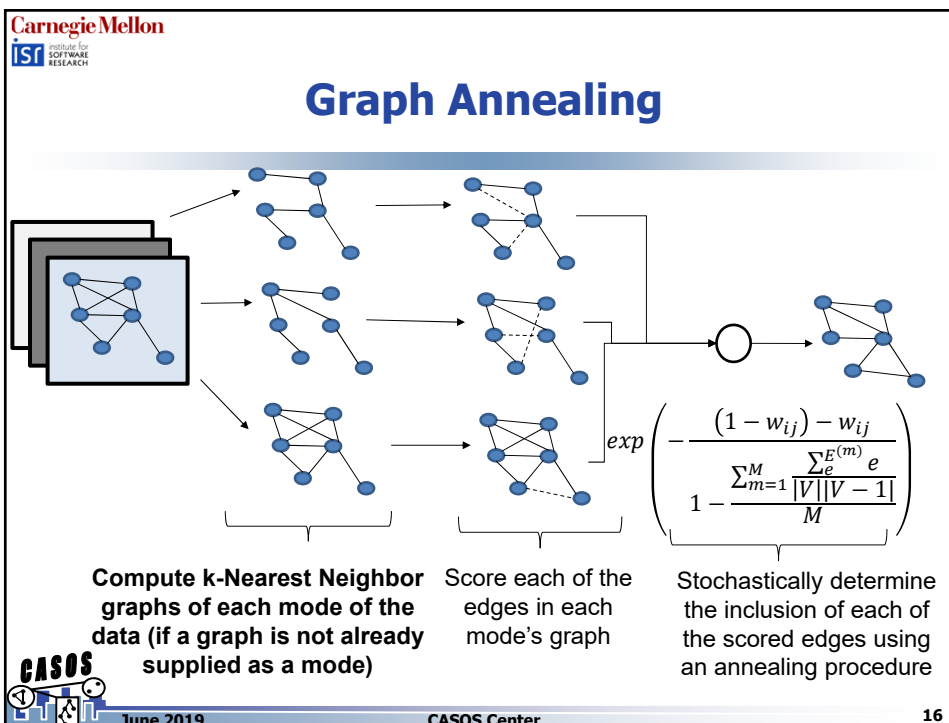
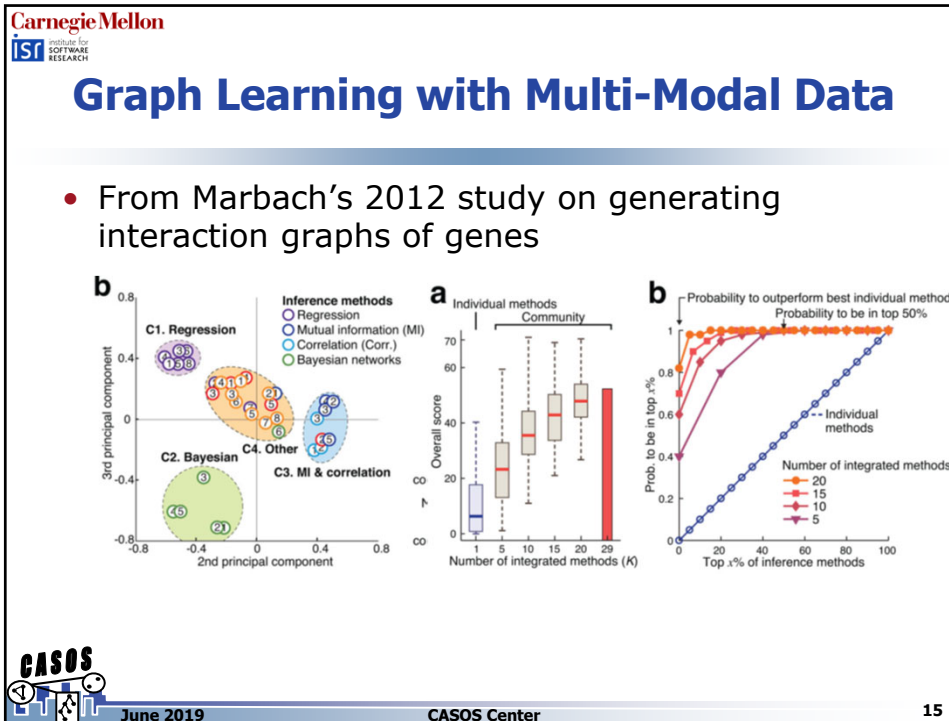
Wang et al. *Similarity network fusion for aggregating data types on a genomic scale*. Nature Methods vol. 11 (3). 2014

- Fusion-by-diffusion
- General idea is to use a matrix diffusion process to combine graphs
 - $P_{t+1}^1 = S^1 \times P_t^2 \times S^{1T}$
 - $P_{t+1}^2 = S^2 \times P_t^1 \times S^{2T}$
 - $G = \frac{P^1 + P^2}{2}$

CASOS

June 2019 CASOS Center 14





Carnegie Mellon
ISI Institute for Software Research

Graph Annealing

$$\exp \left(- \frac{(1 - w_{ij}) - w_{ij}}{1 - \frac{\sum_{m=1}^M \frac{\sum_e^{E(m)} e}{|V||V-1|}}{M}} \right)$$

Compute k-Nearest Neighbor graphs of each mode of the data (if a graph is not already supplied as a mode)

Score each of the edges in each mode's graph

Stochastically determine the inclusion of each of the scored edges using an annealing procedure

CASOS June 2019 CASOS Center 17

Carnegie Mellon
ISI Institute for Software Research

Graph Annealing

$$\exp \left(- \frac{(1 - w_{ij}) - w_{ij}}{1 - \frac{\sum_{m=1}^M \frac{\sum_e^{E(m)} e}{|V||V-1|}}{M}} \right)$$

Compute k-Nearest Neighbor graphs of each mode of the data (if a graph is not already supplied as a mode)

Score each of the edges in each mode's graph

Stochastically determine the inclusion of each of the scored edges using an annealing procedure

CASOS June 2019 CASOS Center 18

Carnegie Mellon


ISI

Institute for
SOFTWARE
RESEARCH

Online News + Misinformation Example

- One of the Kaggle misinformation news datasets
 - ~18,000 articles with ~40/60 'misinformation'/genuine
 - Only preprocessing was to remove articles that did not have text and those with foreign text
- Three potential views
 - The actual words used (Frequency Matrix)
 - The Parts of Speech used (document-by-PoS)
 - Re-publication network (classic network)
- Analyst Questions
 - Is it possible to observe likely areas, or topics of misinformation
 - Are there subgroups within the information? Are certain topics more important than others

CASOS



June 2019

CASOS Center

19

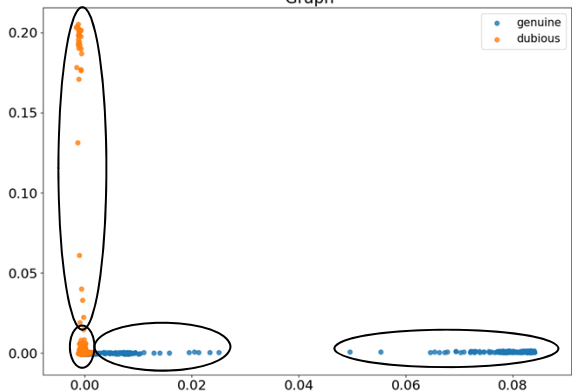
Carnegie Mellon

ISI


Institute for
SOFTWARE
RESEARCH

Visualizing the Fused Data

Spectral Embedding of the Annealed Graph



CASOS

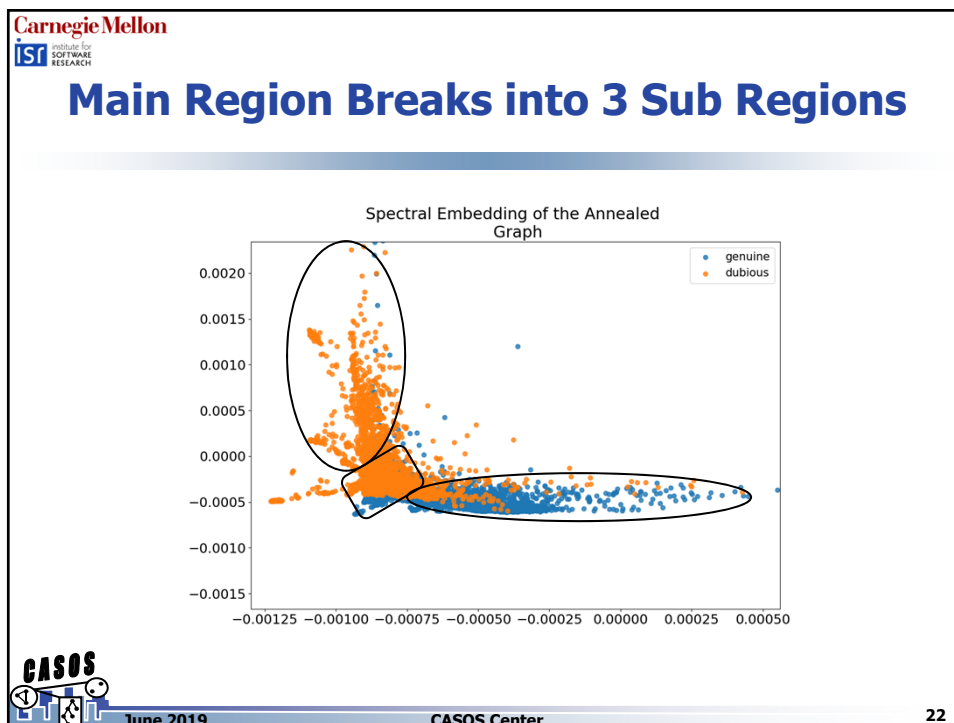
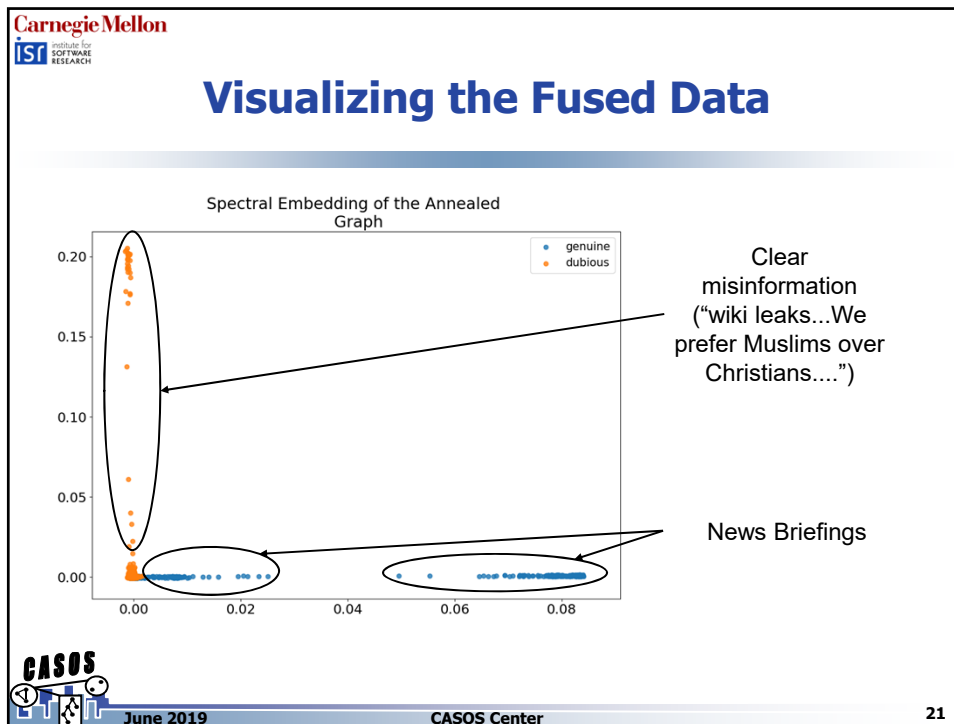



June 2019

CASOS Center

20







Clusters within the Annealed Graph Reflect Topic Groups

- Group 17: War with Russia
 - "Russia Preparing for Nuclear World War III with NATO, USA Breaking News November 5 2016"
- Group 29: News about California
 - "San Diego Struggles to Keep Its Young Tech Talent"
- Group 9: Healthcare and Education (President Obama's legacy)
 - "Republicans 4-Step Plan to Repeal the Affordable Care Act"
 - "NC School District Fights to Keep Pro-Transgender Message in First-Grade Curriculum"
- Group 16: Dakota pipeline, police brutality, oil & gas
 - "With New Study in Hand, Pennsylvanians Reiterate Call for Fracking Ban"
 - "Why Everyone on Facebook Is Checking into Standing Rock, North Dakota"

The Proportions of Misinformation within Each Cluster

Cluster	dubious (red)	genuine (blue)
2	0.68	0.32
3	0.65	0.35
4	0.58	0.42
5	0.48	0.52
6	0.35	0.65
7	0.30	0.70
8	0.12	0.88
9	0.08	0.92
10	0.05	0.95
11	0.02	0.98
12	0.01	0.99
13	0.01	0.99
14	0.01	0.99
15	0.01	0.99
16	0.01	0.99
17	0.01	0.99
18	0.01	0.99
19	0.01	0.99
20	0.01	0.99
21	0.01	0.99



Summary

- Misinformation is difficult to characterize as it has a lot of ways of presenting in the data, based upon many different sociological and psychological theories
 - Using more modes of data results in better supervised prediction of misinformation
- Graph Learning can be used to fuse multi-modal data to help identify misinformation
 - Combine the benefits of the different views with the flexibility of a graph
- Preliminary studies show the fused data model is useful for huge amounts of text-and-network data
 - Aid with understanding the nature of the information