Social Media Analytics

Community Detection
Bi-partite Networks
Cliques and cores

MSBA, 14th Feb, 2022

Dr. Anitesh Barua

David Bruton Jr. Centennial Chair Professor of Business

Distinguished Fellow, INFORMS Information Systems Society

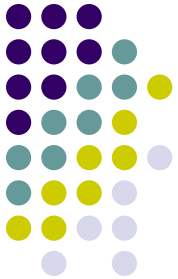
University of Texas Distinguished Teaching Professor

Associate Director, Center for Research in e-Commerce

McCombs School of Business, University of Texas at Austin

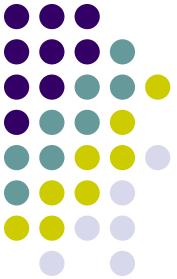
Email: aniteshb@gmail.com

Community Detection: Why Bother?

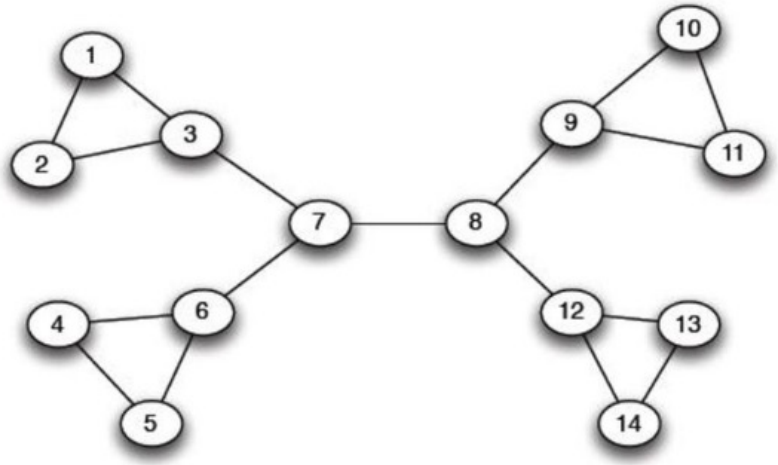


- Detecting networks of fraudulent/rogue websites
  - Many use JavaScript redirects to link to each other to avoid detection through scraping
- Estimating unknown features of users in social networks
- Clustering similar users together
  - Enhance meaningful communication
- Can be a network of products
  - E.g., to show the effect of recommender systems on competition
  - Show that a "community" has products from very different parts of the demand curve
  - https://joshbarua2002.medium.com/who-is-your-competitor-in-the-era-of-the-long-tail-d0ac24fedde8

How to Detect Communities Within Networks



- Common for uni-partite (1-mode) networks
- Girvan-Newman algorithm (divisive algorithm)
- Calculate betweenness centrality of "links"
- How?



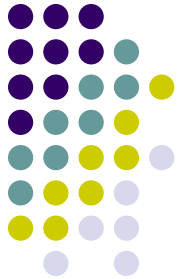
Betweenness(7, 8)= 7x7 = 49

Betweenness(1, 3) = 1X12=12

Betweenness(3, 7) = Betweenness(6, 7) =
Betweenness(8, 9) = Betweenness(8, 12)= 3X11=33

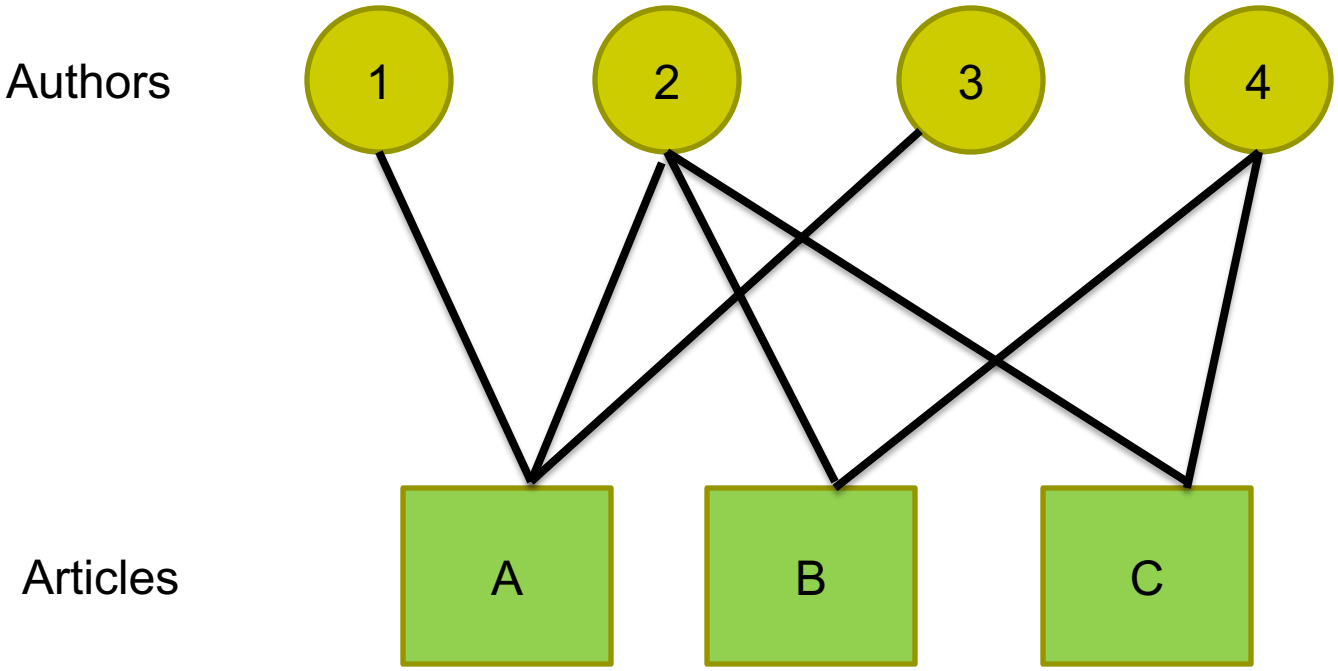
- (i) Cut the link with highest betweenness centrality
- (ii) Recalculate betweenness for all remaining links
- (iii) Cut the link with highest betweenness
- (iv) Repeat (ii) and (iii) until the network disintegrates into disjoint parts
- Excellent article: https://www.analyticsvidhya.com/blog/2020/04/community-detection-graphs-networks/

Types of Networks



- Uni-partite, bi-partite, tri-partite (or multi-partite) networks
- Also called 1-mode, 2-mode, etc.
- Uni-partite: Only one type of nodes (e.g., people)
- Bi-partite: E.g., authors & articles, actors & movies, FB users and their group memberships, etc.

Bi-partite Networks: An Example



No connections between nodes of the same type

Can we reduce this network to 1-mode? How?

2-mode to 1-mode Networks

- 2-mode: Congress(wo)man & age
- How to reduce to 1-mode (person-to-person)?

	Coble	Franks	Goodlatt	Hartzler	McGove	Nadler	Pingree	Polis	Roby	Waters
Coble	0	26	21	29	28	16	24	44	45	7
Franks	26	0	5	3	2	10	2	18	19	19
Goodlatte	21	5	0	8	7	5	3	23	24	14
Hartzler	29	3	8	0	1	13	5	15	16	22
McGovern	28	2	7	1	0	12	4	16	17	21
Nadler	16	10	5	13	12	0	8	28	29	9
Pingree	24	2	3	5	4	8	0	20	21	17
Polis	44	18	23	15	16	28	20	0	1	37
Roby	45	19	24	16	17	29	21	1	0	38
Waters	7	19	14	22	21	9	17	37	38	0

Source: http://www.umasocialmedia.com/socialnetworks/networks-lecture-13-qap-correlation/

Gender & Committees

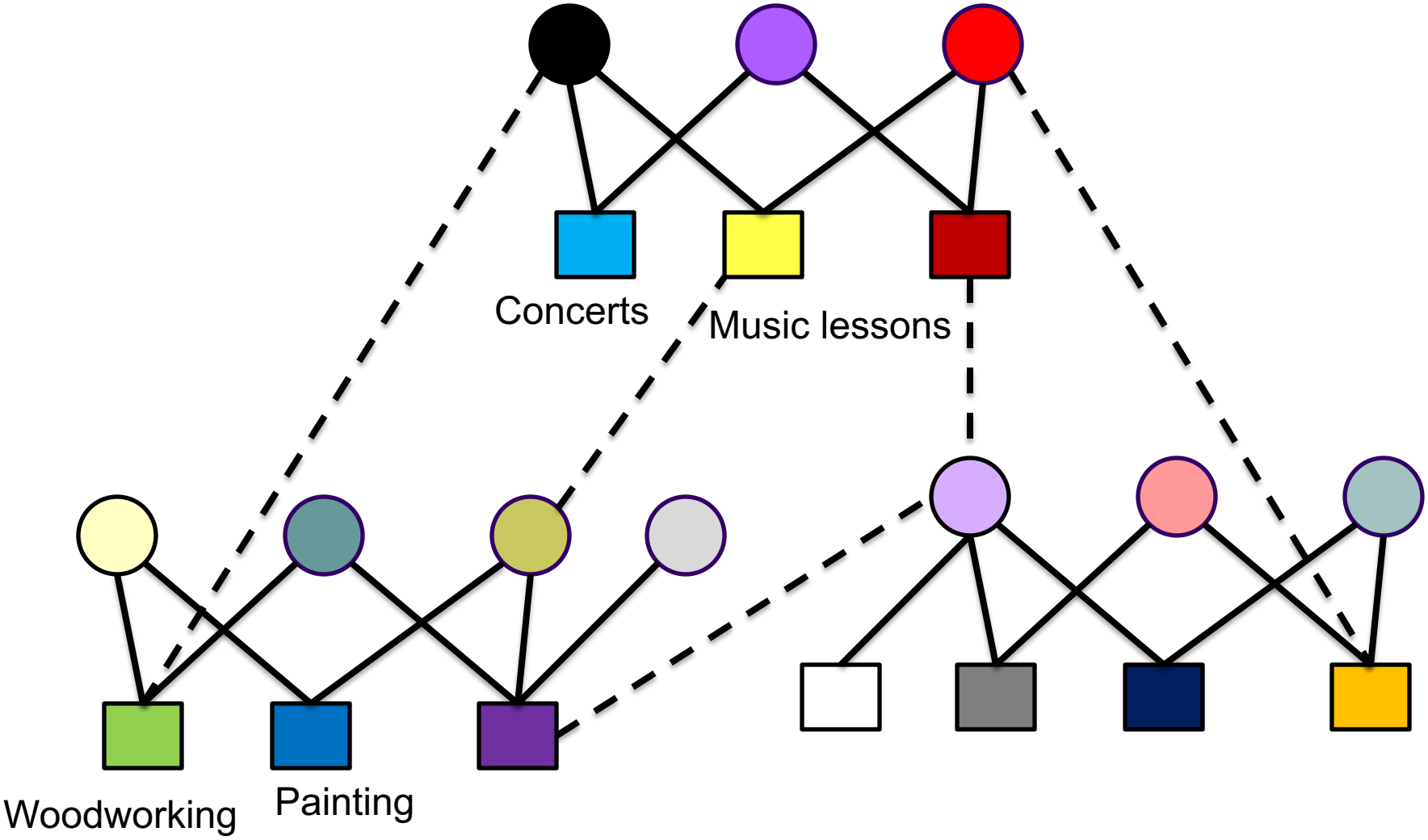
	Coble	Franks	Goodlatt	Hartzler	McGove	Nadler	Pingree	Polis	Roby	Waters
Coble	1	1	1	0	1	1	0	1	0	0
Franks	1	1	1	0	1	1	0	1	0	0
Goodlatte	1	1	1	0	1	1	0	1	0	0
Hartzler	0	0	0	1	0	0	1	0	1	1
McGovern	1	1	1	0	1	1	0	1	0	0
Nadler	1	1	1	0	1	1	0	1	0	0
Pingree	0	0	0	1	0	0	1	0	1	1
Polis	1	1	1	0	1	1	0	1	0	0
Roby	0	0	0	1	0	0	1	0	1	1
Waters	0	0	0	1	0	0	1	0	1	1

	Coble	Franks	Goodlatt	Hartzler	McGove	Nadler	Pingree	Polis	Roby	Waters
Coble	1	1	1	0	0	1	0	1	0	1
Franks	1	2	2	1	0	1	1	1	1	1
Goodlatte	1	2	2	1	0	1	1	1	1	1
Hartzler	0	1	1	2	1	0	2	0	2	0
McGovern	0	0	0	1	2	0	1	1	1	0
Nadler	1	1	1	0	0	1	0	1	0	1
Pingree	0	1	1	2	1	0	2	0	2	0
Polis	1	1	1	0	1	1	0	2	0	1
Roby	0	1	1	2	1	0	2	0	2	0
Waters	1	1	1	0	0	1	0	1	0	1

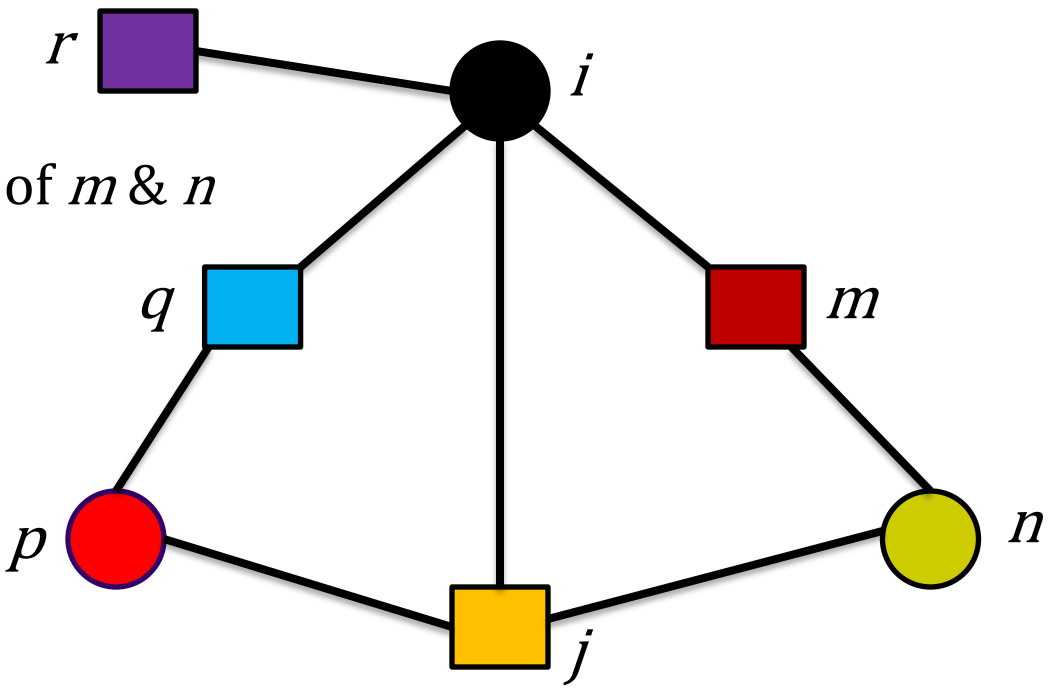
Source: http://www.umasocialmedia.com/socialnetworks/networks-lecture-13-qap-correlation/

Communities in Bi-partite Networks

- Densely linked parts of a bi-partite network constitute communities
- E.g., people's memberships in a set of activities



Detecting Communities in Bi-partite Networks



k_m & k_n degrees of m & n

For a pair of nodes, *i* and *j*, let *m* and *n* be neighbors of *i* and *j* respectively

q_{ijmn} = 1 if *m* and *n* are connected, 0 otherwise. θ_{ijmn} has the opposite definition.

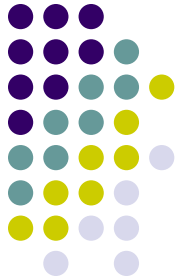
Edge clustering coefficient $C(i,j) =$

squares that currently include *i-j* / possible # squares that include *i-j*

$$\frac{\sum_{m=1}^{k_i}\sum_{n=1}^{k_j} q_{ijmn}}{\sum_{m=1}^{k_i}\sum_{n=1}^{k_j} \theta_{ijmn} + \sum_{m=1}^{k_i}(k_m - 1) + \sum_{n=1}^{k_j}(k_n - 1) - \sum_{m=1}^{k_i}\sum_{n=1}^{k_j} q_{ijmn}}$$

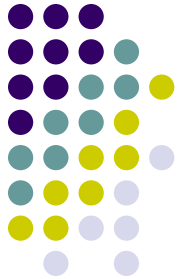
Source: P. Zhang et al. 2013

Detecting Communities in Bi-Partite Networks



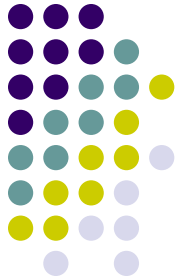
- Start dropping links with smallest clustering coefficient
- Network will start splitting up
- But much easier to first divide into two 1-mode networks
- The apply G-N or other algorithms to detect communities within 1-mode networks
- We do lose some information
- Example: https://medium.com/@adibarua2002/creating-smarter-online-communities-with-nlp-and-network-analytics-147810d3cee5?source=friends_link&sk=d7f5809dfa0e7cc774448615e9287de5

Sample Project on Bi-Partite Networks



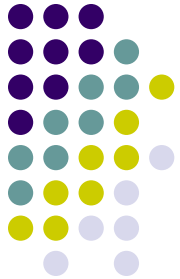
- Readers and books
- How to create a manageable network
- Compare recommendations based on
  - Book-to-book similarity (uni-partite)
  - Person-to-person and then recommending books that similar readers have read but not the focal reader.
- Does the second method provide more variety?

Identifying Trolls, Spammers and Fake Content Creators



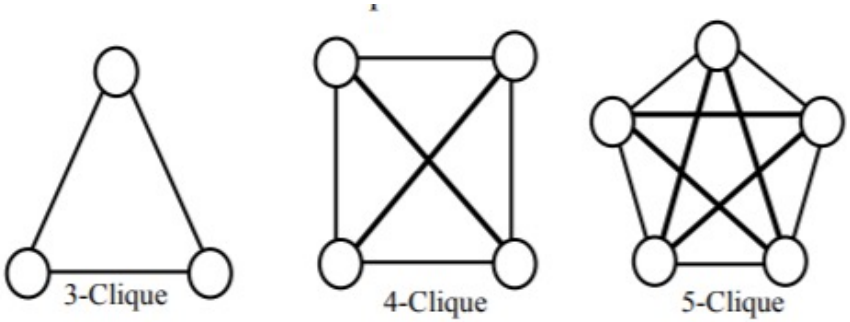
- Trolls, spammers, bots
- Creators of fake content
- Is there anything common across them?
- Was not too difficult to detect fake content (e.g., reviews) from content in the past
- Why is it difficult to detect now?
- Network analytics may help

They Come in Many Flavors



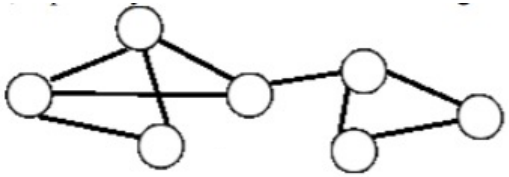
- ~ 19 million bot accounts tweeted in support of either Trump or Clinton in 2016
- > 1,000 paid Russian trolls spread “fake news” on Hillary Clinton
- CNN mobile app received 100s of thousands of 1-star reviews after the network’s treatment of a certain Reddit user
- The Boca Raton Resort hotel got a huge number of negative reviews (1000s) after a Youtube star angry at his treatment rallied his fanbase to retaliate online
- **<u>But may be more connected than regular users!</u>**

Of “Cliques” and “Cores”



- Definition of an n-clique?
- What is the significance of a clique?
- A large network will consist of many cliques
- Often the largest clique is of interest

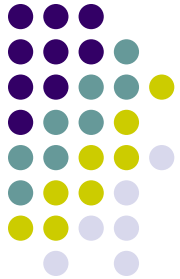
K-core



A 2-core

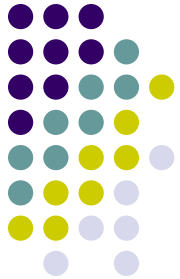
- Definition of a k-core?
- What is the significance?
- Largest k for a network is of special interest

Hypotheses (Can be Tested)



- Spammers have larger k-cores than non-spammers
- Spammers have larger n-cliques than non-spammers
- Spammers have higher network density

Calculating Cliques and k-cores



- Lots of python code available on GitHub and other sites
- https://www.kaggle.com/mayeesha/network-analysis-for-dummies-stackoverflow-data
- https://s3.amazonaws.com/assets.datacamp.com/production/course_3286/slides/ch3_slides.pdf
- https://towardsdatascience.com/intro-to-graphs-in-python-using-networkx-cfc84d1df31f
- Maximum value of k in k-cores within a network
  - https://github.com/chibuta/k-core-subgraph

Data Issues

- If collecting primary data
  - Twitter → Find out important hashtags (e.g., anti- vs. pro-vaxxer, anti-climate change vs. pro)

Stance	Hashtags
Pro-vaccination	*VaccinesSaveLives, VaccinesWork, WorldImmunizationWeek, VaxWithMe, HealthForAll, WiW, ThankYouLaura*
Anti-vaccination	*LearnTheRisk, VaccineInjury, VaccineDeath, VaccineDamage, VaccinesCauseAutism, CDCFraud, CDCWhistleBlower, CDCTruth, WakeUpAmerica, HearUs, HealthFreedom*
Unidentified	*Vaccine, Vaccines, Vaccinate, VaccinateUS*

  - Other sources (like discussion forum): Separate source and target during scraping.
    - E.g., creator of an original post in one column
    - Person (or people) commenting in another column
- Many useful sources of archived network data
  - https://github.com/benedekrozemberczki/datasets
  - Other sources on GitHub, some on Kaggle