COMMUNITY DETECTION IN SOCIAL NETWORKS

LEARNING PYTHON THE HARD WAY

MINING SOCIAL MEDIA DATA – SESSION 4

APRIL 4, 2017

Toronto Data Literacy Group

TWG

KΣ SOLUTIONS

WHAT WE HAVE COVERED SO FAR?

- Social Media Data Collection
  - API
  - Web Scraping
  - Public Available Datasets
- Social Media Data Analytics
  - Text Mining
  - Social Network Analysis
- Python Programming Basics
- Everything We Learned So Far Can Be Found Here:
  - https://github.com/cindyzhong/trt_data_lit_grp_python
  - https://github.com/m3rl10n/PyLadies_ScrapingTalk

KEY IDEAS FROM OUR LAST SESSION

- Components of Network
  - Nodes
  - Edges
- Centrality Analysis – Which node is important in an network?
  - Degree Centrality (Normalized Degree Centrality)
    - Importance of a node is determined by the number of nodes adjacent to it
    - Fraction of nodes node Vj is connected to

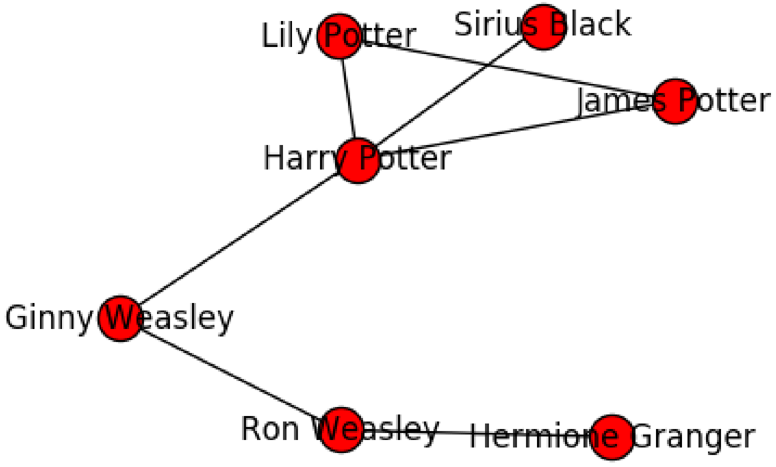
$$C_D(v_i) = d_i = \sum_j A_{ij}$$

  - Closeness Centrality
    - If a person is very close to all other people in the network, he/she should be central of the network
    - higher values of closeness indicate higher centrality

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)},$$

  - betweenness Centrality
    - Node betweenness: number of shortest path that pass through the node
    - A person is important if he/she lies on the critical communication path

$$c_B(v) = \sum_{s,t\in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$



For example, for Ginny Weasley in the graph above,
Cd = 2/(7-1) = 0.33
Cc = (7-1)/(1+1+2+2+2+2) = 0.6

KEY IDEAS FROM OUR LAST SESSION

- Comparing Centrality Measures
  - Centrality measures usually correlates with each other
  - When there is difference, we can find interesting individuals

	Low Cd	Low Cc	Low Cb
High Cd		Leader of sub community	Well connected but redundant
High Cc	Key player Hide behind active node		Near many people but so are others
High Cb	Boundary Spanner Few connections Crucial to keep network together	Monopolize communication ties from a small # of people	

WHAT IS A COMMUNITY?

- Can also be called groups, subgroups, clusters, or modules
  - A substructure (subset of vertices) of a graph with dense linkage between the members of the community and sparse density outside the community
- Community detection
  - discovering groups in a network where individuals' group memberships are not explicitly given
- Many interesting aspects to study about, for example,
  - Social phenomena
  - Data compression (Node -> Community)
  - Temporal development of network, link prediction
  - Collaborative filtering

WHAT IS A COMMUNITY?

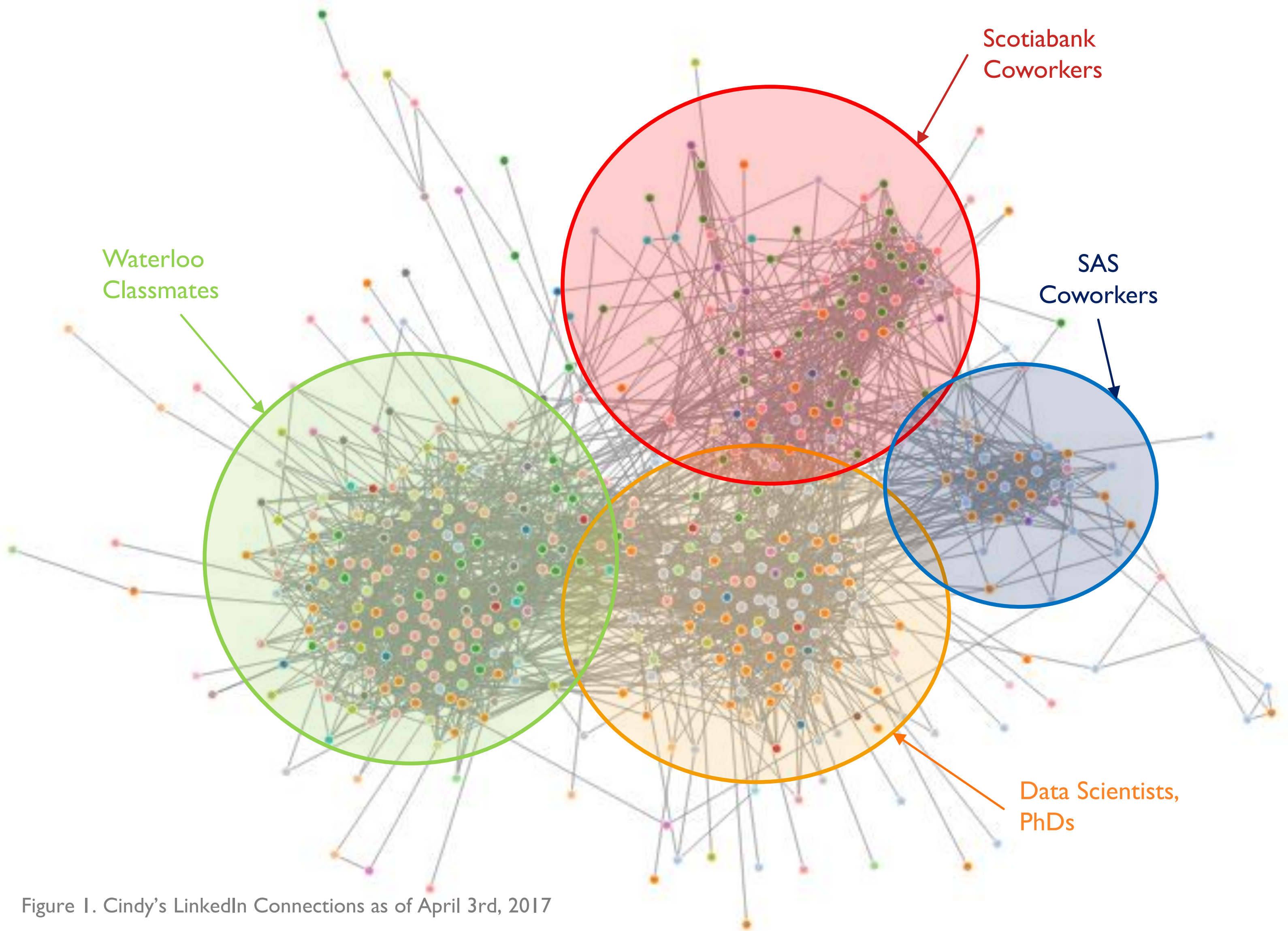


Figure 1. Cindy's LinkedIn Connections as of April 3rd, 2017

Try it Yourself

https://github.com/craigtutterow/sna-js

MODULARITY OF A NETWORK

- Measures the strength of a community partition by taking into account the degree distribution
- Given a network of *n* nodes and *m* edges, the *expected number of edges* between nodes v_i and v_j is $d_i d_j/2m$, where d_i and d_j are the degrees of node v_i and v_j, respectively.
- Strength of community effect
  - how far the true network interaction between nodes *i* and *j* (A_{ij}) deviates from the expected random connections
- Modularity:

$$Q = \frac{1}{2m} \sum_{\ell=1}^{k} \sum_{i \in C_\ell, j \in C_\ell} \left(A_{ij} - d_i d_j / 2m \right)$$

- A value between -1 and 1
- A larger value indicates a good community structure
  - dense connections between the nodes within modules
  - sparse connections between nodes in different module

EDGE BETWEENNESS

- Edge Betweenness Algorithm
  - Proposed by Girvan & Newman
  - Divisive Hierarchical Clustering
  - Edge betweenness : number of shortest paths that pass along one edge
  - Edges connecting different groups are more likely to be contained in multiple shortest paths simply because in many cases they are the only option to go from one group to another. For example, e(NL,LV) and e(HG,LV)
- The edge with higher betweenness tends to be the bridge between two communities.
- The underlying community structure of the network will be much clear after removing edges with high edge betweenness.

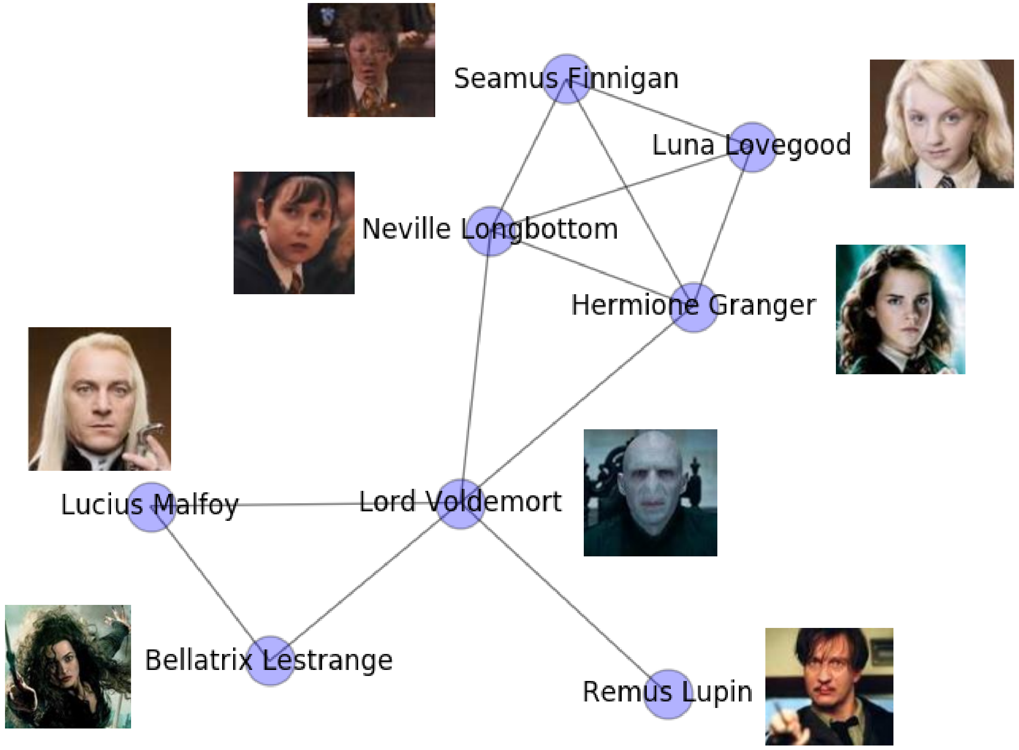
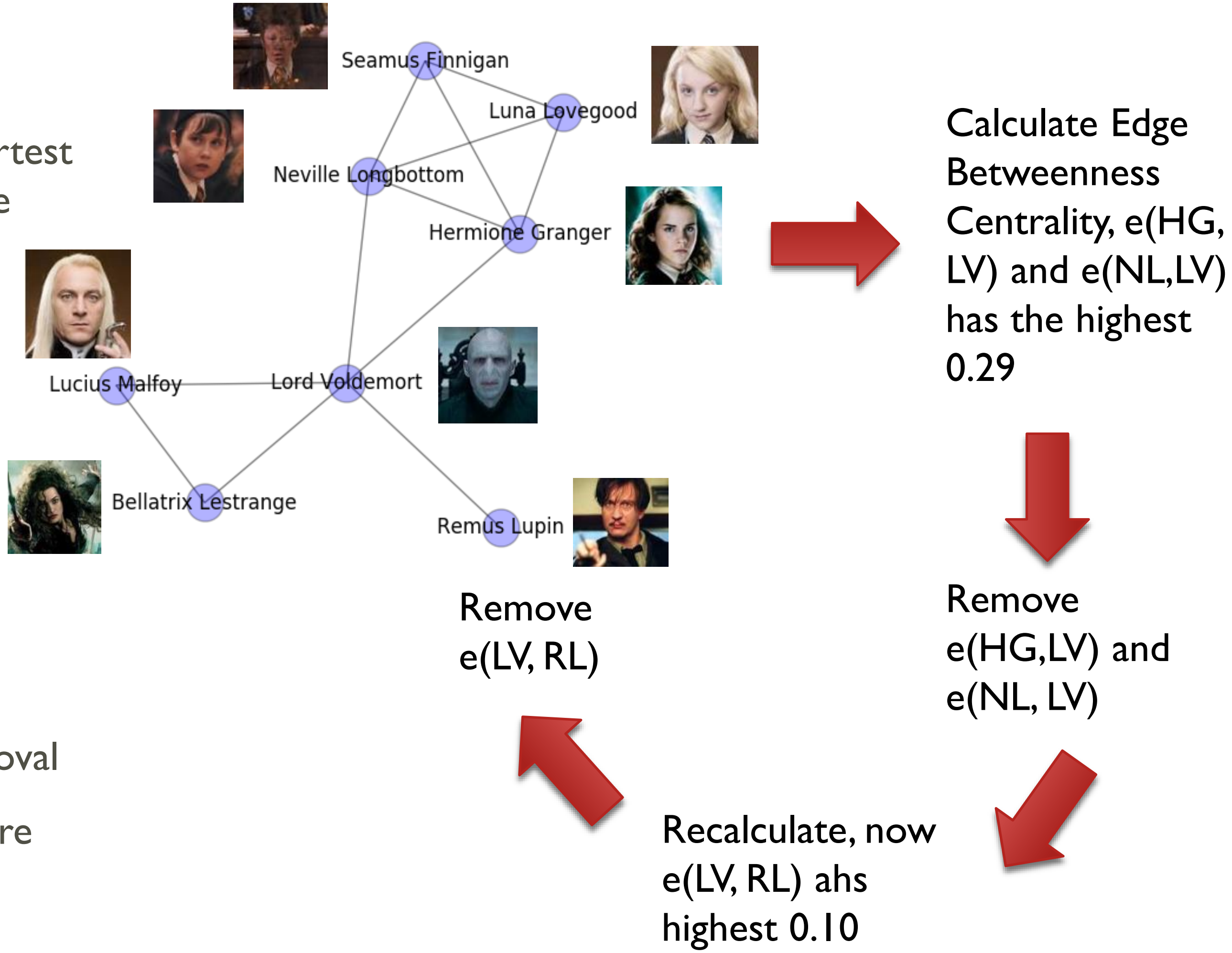


Figure 2. A subgraph of several Harry Potter Characters

EDGE BETWEENNESS

- Basic implementation:
  1. Calculate the edge betweenness (the number of shortest paths pass along one edge) of all existing edges in the network
  2. Remove edge with highest betweenness
  3. Recalculate edge betweenness for all edges
  4. Remove step 2,3 until no edges remain

- Problem with the implementation:
  - Computationally intensive because the betweenness scores have to be re-calculated after every edge removal
  - Result is basically a dendrogram, no guidance on where to cut the dendrogram to obtain the final groups



Calculate Edge Betweenness Centrality, e(HG, LV) and e(NL,LV) has the highest 0.29

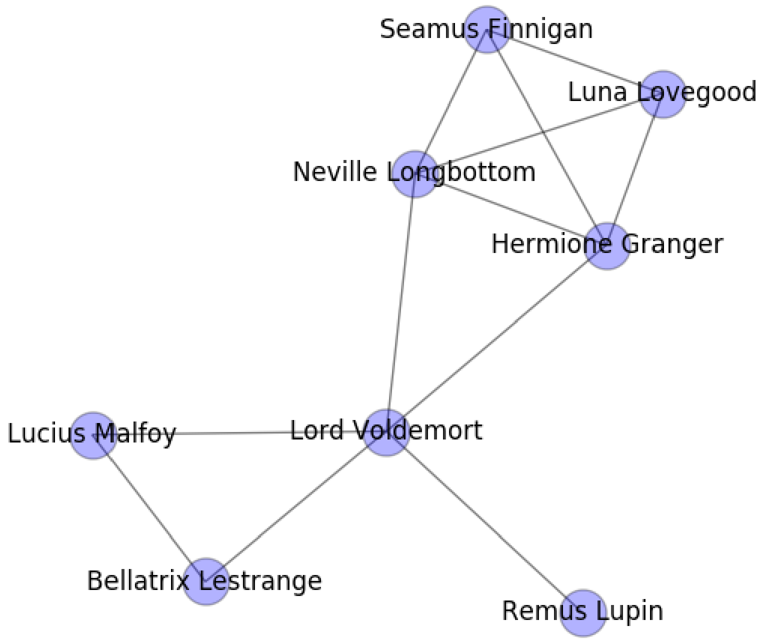
Remove e(HG,LV) and e(NL, LV)

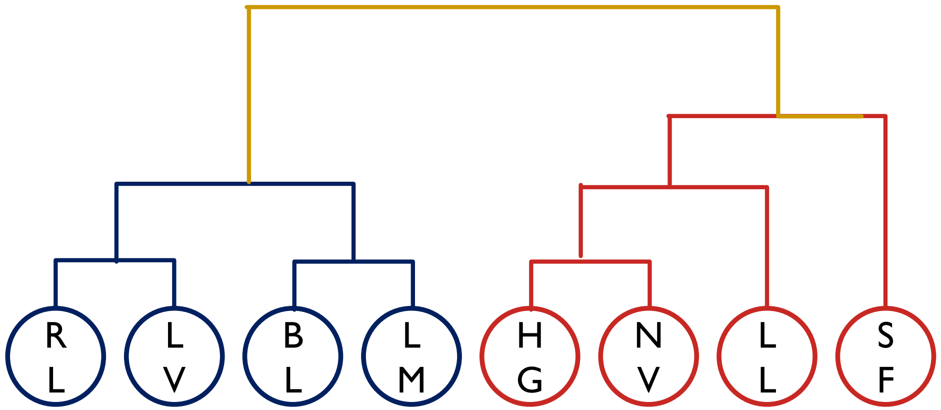
Recalculate, now e(LV, RL) ahs highest 0.10

Remove e(LV, RL)

LOUVAIN METHOD

- Proposed by Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre
- Fast Greedy Method
- Agglomerative Hierarchical Clustering
- Basic Implementation:
  1. Initialize each node as a singleton community
  2. Calculate the expected improvement of modularity for each pair of communities,
  3. Choose a community pair that gives the maximum improvement of modularity and merges them into a new community
  4. Repeated until no community pairs merge leads to an increase in modularity
- Very fast O(n log n) for most sparse, hierarchical, networks


Seamus Finnigan
Luna Lovegood
Neville Longbottom
Hermione Granger
Lucius Malfoy
Lord Voldemort
Bellatrix Lestrange
Remus Lupin


R L
L V
B L
L M
H G
N V
L L
S F