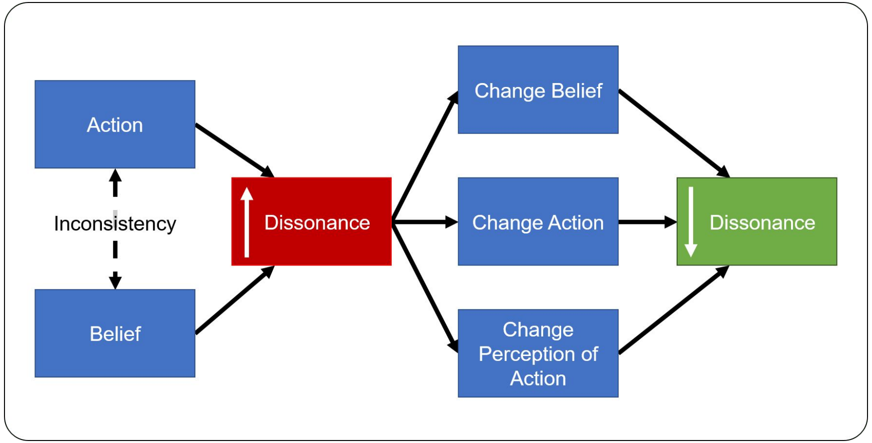
# Social network Analysis

A short overview

Riccardo La Grassa

# Introduction

Tendency to connect with another entities




Six degrees of separation

| Year | Distance | |
|---|---|---|
| 2008 | 5.28 | |
| 2011 | 4.74 | |
| 2016 | 4.57 | |
| Distances as reported in Feb 2016 [38][41] | | |

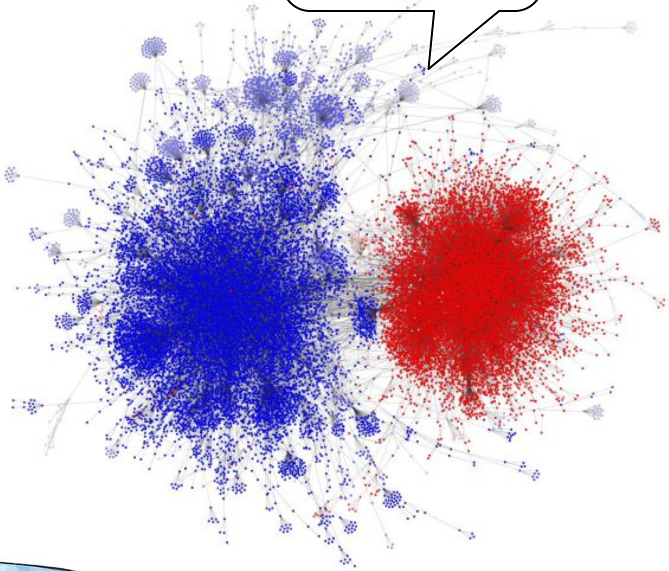
Facebook

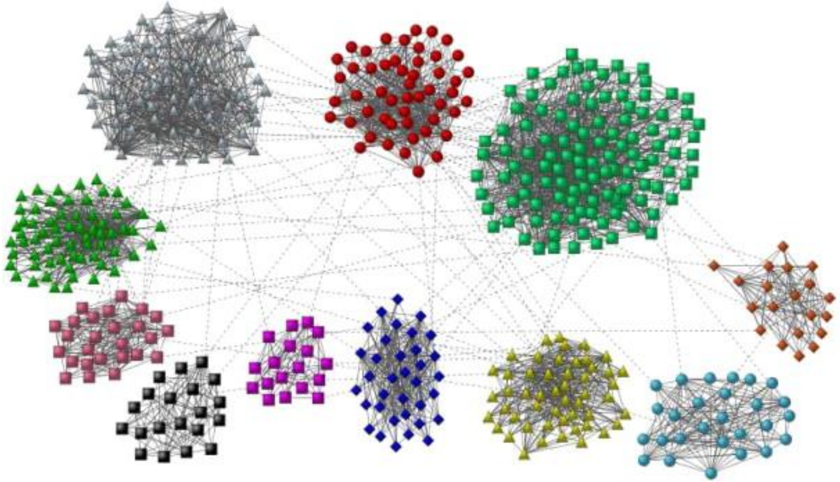
*“SNA is the process of investigating social structures through the use of networks and graph theory”*

Otte, Evelien; Rousseau, Ronald (2002). "Social network analysis: a powerful strategy, also for the information sciences". *Journal of Information Science*. **28** (6): 441–453.

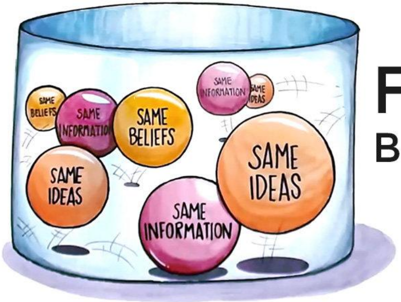
# Consequences

Polarization





Echo-chamber

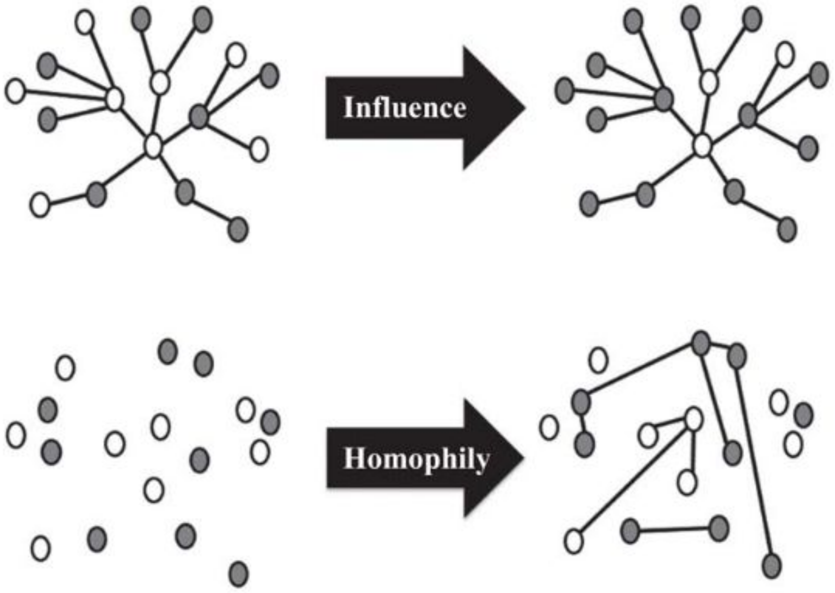



Filter
Bubbles

# Homophily and Influence

Influence: the capacity to have an effect on the character, development, or behaviour of someone or something, or the effect itself

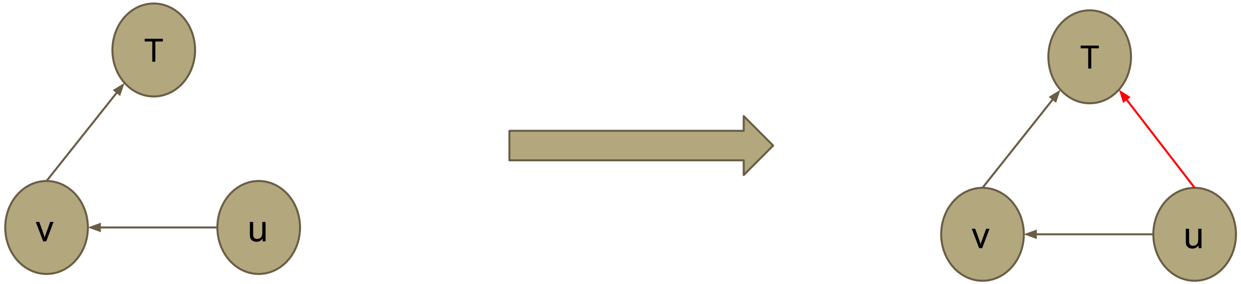
Homophily: people tend to be similar to their friends



# Triadic closure

$$\text{t(0)=}\left\{(u,v,target)\;\middle|\;(u,v)\in E_a\wedge(v,target)\in E_a\wedge(u,target)\notin E_a,\;\;u\neq target\right\}\tag{14}$$

$$\text{t(n)=}\left\{(u,v,target)\;\middle|\;(u,v)\in E_a\wedge(v,target)\in E_a\wedge(u,target)\in E_a,\;\;u\neq target\right\}\tag{15}$$



# Centrality measures

Not all nodes are equally important

Centrality Analysis -> Discover the most important nodes in a network

Most metrics used:

- Degree Centrality
- Closeness Centrality
- Betweenness Centrality

# Degree centrality

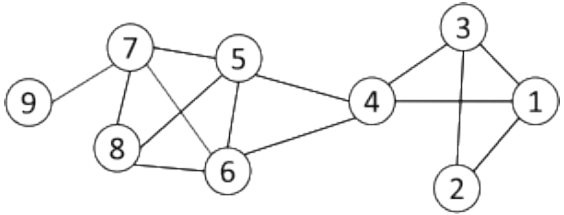
The importance of a node is determined by the number of nodes adjacent to it

Degree Centrality

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

Norm degree centrality

$$C'_D(v_i) = d_i/(n-1)$$




For node 1, degree centrality is 3;
Normalized degree centrality is
3/(9-1)=3/8.

# Closeness Centrality

"Central" nodes are important, as they can reach the whole network more quickly than non-central nodes

Closeness Centrality

$$C_C(v_i) = \left[ \frac{1}{n-1} \sum_{j \neq i}^{n} g(v_i, v_j) \right]^{-1} = \frac{n-1}{\sum_{j \neq i}^{n} g(v_i, v_j)}$$

# Example

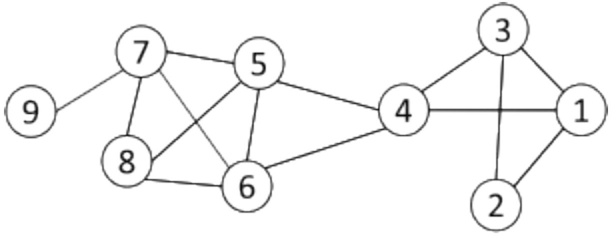



Table 2.1: Pairwise geodesic distance

| Node | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 1 | 2 | 2 | 3 | 3 | 4 |
| 2 | 1 | 0 | 1 | 2 | 3 | 3 | 4 | 4 | 5 |
| 3 | 1 | 1 | 0 | 1 | 2 | 2 | 3 | 3 | 4 |
| 4 | 1 | 2 | 1 | 0 | 1 | 1 | 2 | 2 | 3 |
| 5 | 2 | 3 | 2 | 1 | 0 | 1 | 1 | 1 | 2 |
| 6 | 2 | 3 | 2 | 1 | 1 | 0 | 1 | 1 | 2 |
| 7 | 3 | 4 | 3 | 2 | 1 | 1 | 0 | 1 | 1 |
| 8 | 3 | 4 | 3 | 2 | 1 | 1 | 1 | 0 | 2 |
| 9 | 4 | 5 | 4 | 3 | 2 | 2 | 1 | 2 | 0 |

$$C_C(3) = \frac{9-1}{1+1+1+2+2+3+3+4} = 8/17 = 0.47,$$

$$C_C(4) = \frac{9-1}{1+2+1+1+1+2+2+3} = 8/13 = 0.62.$$

# Betweenness centrality

Nodes with high betweenness are important in communication and information diffusion

$$C_B(v_i) = \sum_{v_s \neq v_i \neq v_t \in V, s<t} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

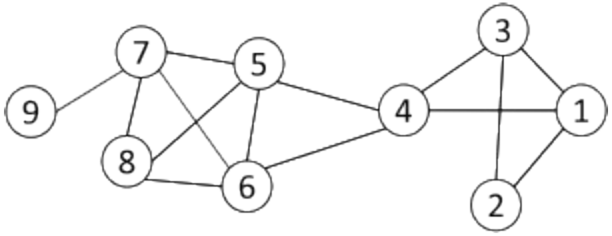
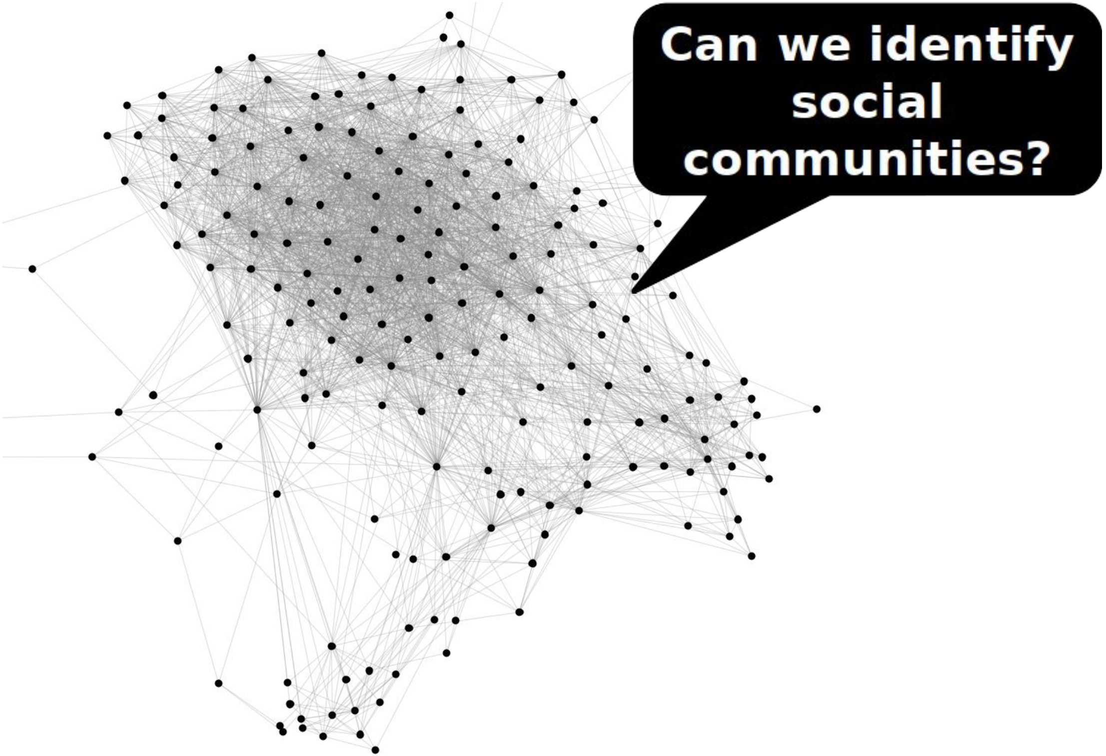



Table 2.2: $\sigma_{st}(4)/\sigma_{st}$

| | $s = 1$ | $s = 2$ | $s = 3$ |
|---|---|---|---|
| $t = 5$ | 1/1 | 2/2 | 1/1 |
| $t = 6$ | 1/1 | 2/2 | 1/1 |
| $t = 7$ | 2/2 | 4/4 | 2/2 |
| $t = 8$ | 2/2 | 4/4 | 2/2 |
| $t = 9$ | 2/2 | 4/4 | 2/2 |

$$C_B(4) = 15$$

# Community detection



# Idea of KL Algorithm

Start with any initial partition X and Y.

A pass or iteration means exchanging each vertex A  X with each vertex B  Y exactly once:

1. For i := 1 to n do

    From the unlocked (unexchanged) vertices,

    choose a pair (A,B) s.t. gain(A,B) is largest.

    Exchange A and B. Lock A and B.

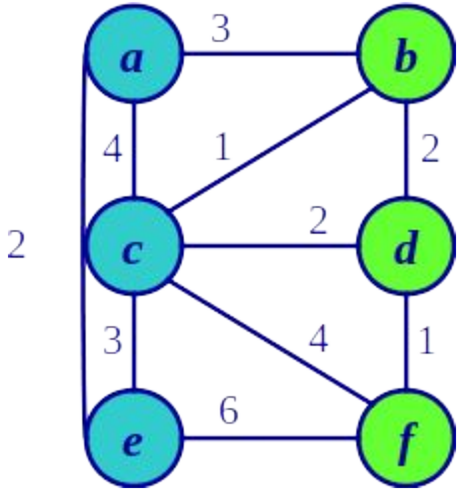
    Let gi = gain(A,B).

2. Find the k s.t. G=g1+...+gk is maximized.

3. Switch the first k pairs.

Repeat the pass until there is no improvement (G=0).

# KL Algorithm 1




Given:

Initial weighted graph G with

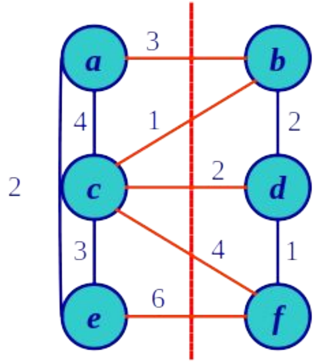
V(G) = { a, b, c, d, e, f }

Start with two partition with equal size

X = { a, c, e }

Y = { b, d, f }

# KL Algorithm 2




cut-size = 3+1+2+4+6 = 16

Compute the gain values of moving node x to the others set:

$G_x = E_x - I_x$

$E_x$ = cost of edges connecting node x with the other group (extra)

$I_x$ = cost of edges connecting node x within its own group (intra)

$X = \{ a, c, e \}$

$Y = \{ b, d, f \}$

$G_a = E_a - I_a = -3 \quad (= 3 - 4 - 2)$

$G_c = E_c - I_c = \quad 0 \quad (= 1 + 2 + 4 - 4 - 3)$

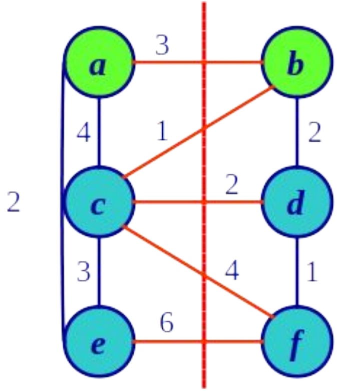
$G_e = E_e - I_e = +1 \quad (= 6 - 2 - 3)$

$G_b = E_b - I_b = +2 \quad (= 3 + 1 - 2)$

$G_d = E_d - I_d = -1 \quad (= 2 - 2 - 1)$

$G_f = E_f - I_f = +9 \quad (= 4 + 6 - 1)$

# KL Algorithm 3




Cost saving when exchanging $a$ and $b$ is essentially $G_a + G_b$

However, the cost saving 3 of the direct edge was counted twice. But this edge still connects the two groups

Hence, the real "gain" (i.e. cost saving) of this exchange is $g_{ab} = G_a + G_b - 2c_{ab}$

$X = \{ a, c, e \}$

$Y = \{ b, d, f \}$

$G_a = E_a - I_a = -3 \ \ (= 3 - 4 - 2)$

$G_b = E_b - I_b = +2 \ \ (= 3 + 1 - 2)$

$g_{ab} = G_a + G_b - 2c_{ab} = -7 \ (= -3 + 2 - 2 \cdot 3)$

# KL Algorithm 4

**Compute all the gains**

$$g_{ab} = G_a + G_b - 2w_{ab} = -3 + 2 - 2 \cdot 3 = -7$$

$$g_{ad} = G_a + G_d - 2w_{ad} = -3 - 1 - 2 \cdot 0 = -4$$

$$g_{af} = G_a + G_f - 2w_{af} = -3 + 9 - 2 \cdot 0 = +6$$

$$g_{cb} = G_c + G_b - 2w_{cb} = 0 + 2 - 2 \cdot 1 = 0$$

$$g_{cd} = G_c + G_d - 2w_{cd} = 0 - 1 - 2 \cdot 2 = -5$$

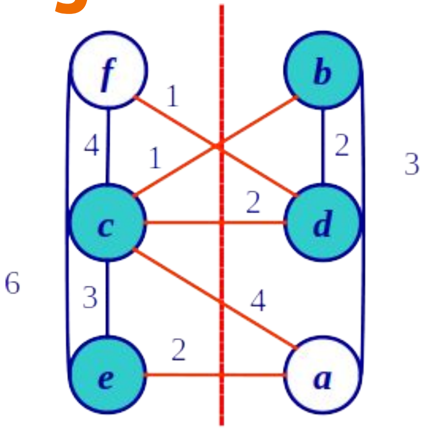
$$g_{cf} = G_c + G_f - 2w_{cf} = 0 + 9 - 2 \cdot 4 = +1$$

$$g_{eb} = G_e + G_b - 2w_{eb} = +1 + 2 - 2 \cdot 0 = +3$$

$$g_{ed} = G_e + G_d - 2w_{ed} = +1 - 1 - 2 \cdot 0 = 0$$

$$g_{ef} = G_e + G_f - 2w_{ef} = +1 + 9 - 2 \cdot 6 = -2$$

# KL Algorithm 5



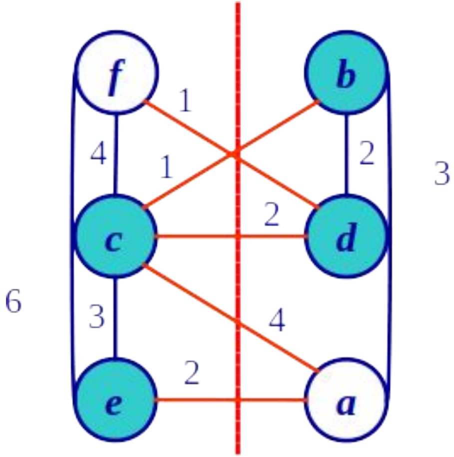

cut-size = 10

$X' = \{ c, e \}$
$Y' = \{ b, d \}$

**Update the G-values of unlocked nodes**

$$G'_c = G_c + 2c_{ca} - 2c_{cf} = 0 + 2(4 - 4) = 0$$
$$G'_e = G_e + 2c_{ea} - 2c_{ef} = 1 + 2(2 - 6) = -7$$
$$G'_b = G_b + 2c_{bf} - 2c_{ba} = 2 + 2(0 - 3) = -4$$
$$G'_d = G_d + 2c_{df} - 2c_{da} = -1 + 2(1 - 0) = 1$$

# KL Algorithm 6




cut-size = 10

$X' = \{ c, e \}$

$Y' = \{ b, d \}$

**Compute the gains**

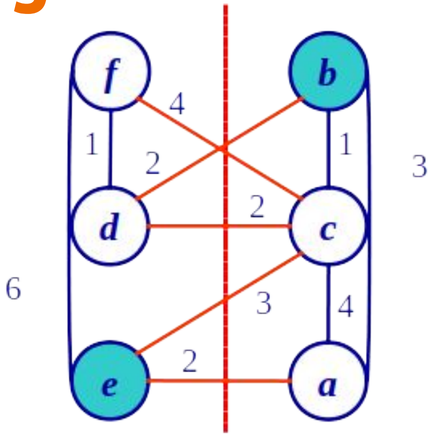
$$g'_{cb} = G'_c + G'_b - 2c_{cb} = 0 - 4 - 2 \cdot 1 = -6$$

$$\boxed{g'_{cd} = G'_c + G'_d - 2c_{cd} = 0 + 1 - 2 \cdot 2 = -3}$$

$$g'_{eb} = G'_e + G'_b - 2c_{eb} = -7 - 4 - 2 \cdot 0 = -11$$

$$g'_{ed} = G'_e + G'_d - 2c_{ed} = -7 + 1 - 2 \cdot 0 = -6$$

# KL Algorithm 7




cut-size = 13

$X'' = \{ e \}$
$Y'' = \{ b \}$

Update the G-values of unlocked nodes

$$G''_e = G'_e + 2c_{ed} - 2c_{ec} = -7 + 2(0 - 3) = -1$$
$$G''_b = G'_b + 2c_{bd} - 2c_{bc} = -4 + 2(2 - 1) = -2$$

Compute the gains

Pair with max. gain is $(e, b)$

$$g''_{eb} = G''_e + G''_b - 2c_{eb} = -1 - 2 - 2 \cdot 0 = -3$$

# Complexity

$O(n^2)$ time to find the best pair to exchange.

n pairs exchanged.

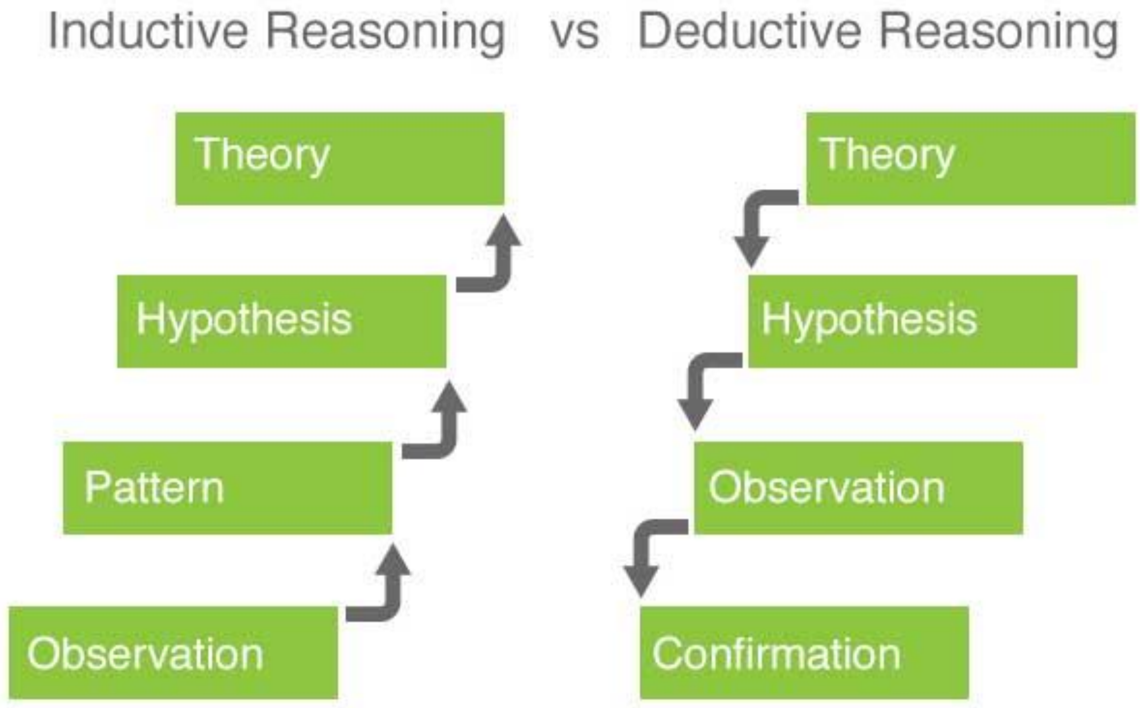
Total time is $O(n^3)$ per pass.

# Collective Classification

Iterative Classification algorithm (ICA)

**Algorithm** $ICA$(Graph $G = (N, A)$, Weights: $[w_{ij}]$, Node Class Labels: $\mathcal{C}$,
Base Classifier: $\mathcal{A}$, Number of Iterations: $T$)
**begin**
**repeat**
Extract link features at each node with current training data;
Train classifier $\mathcal{A}$ using both link and content features of current training data and predict labels of test nodes;
Make (predicted) labels of most "certain" $n_t/T$ test nodes final, and add these nodes to training data, while removing them from test data;
**until** $T$ iterations;
**end**

# Reasoning methods

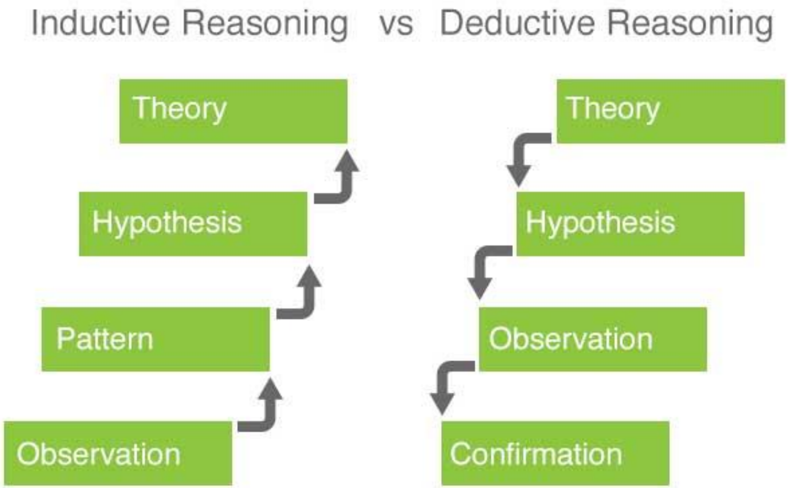


# Social network Analysis

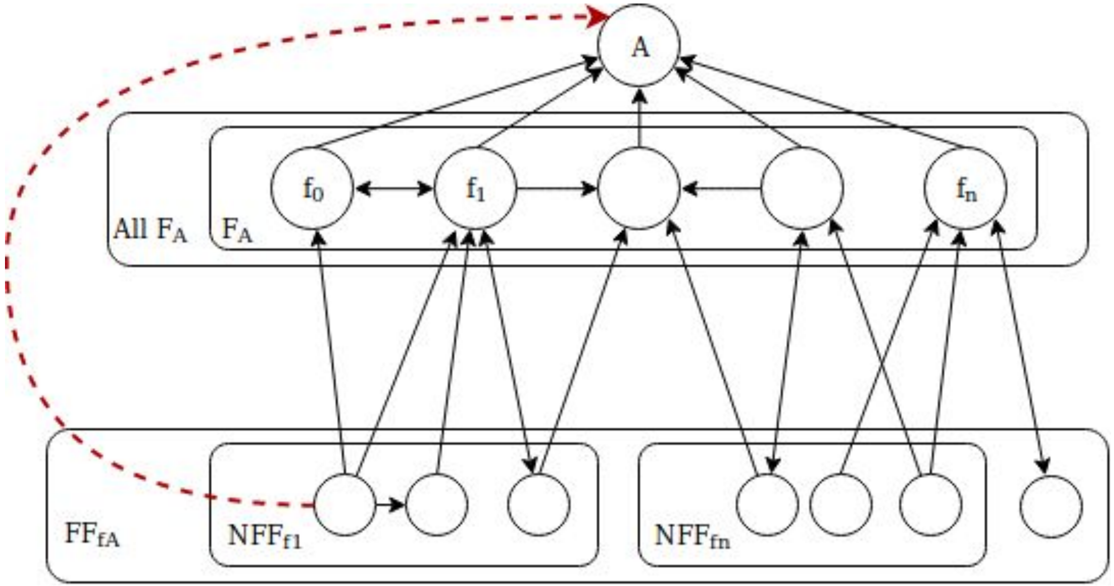
1. Data collection system (DCS)
2. Data analysis (Pattern extraction)
3. Hypothesis
4. Theory

---

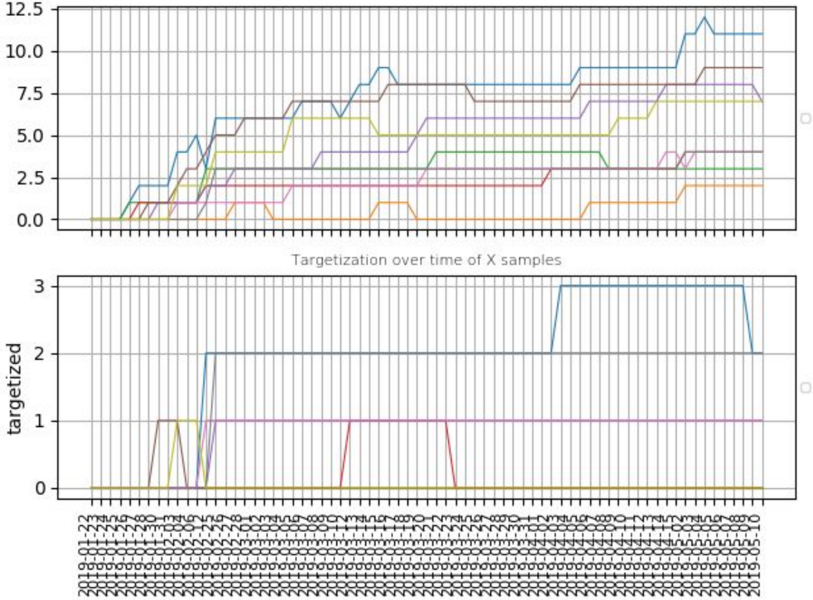
1. Theory
2. Hypothesis
3. Observation (DCS, Data analysis)
4. Confirmation (Proof)

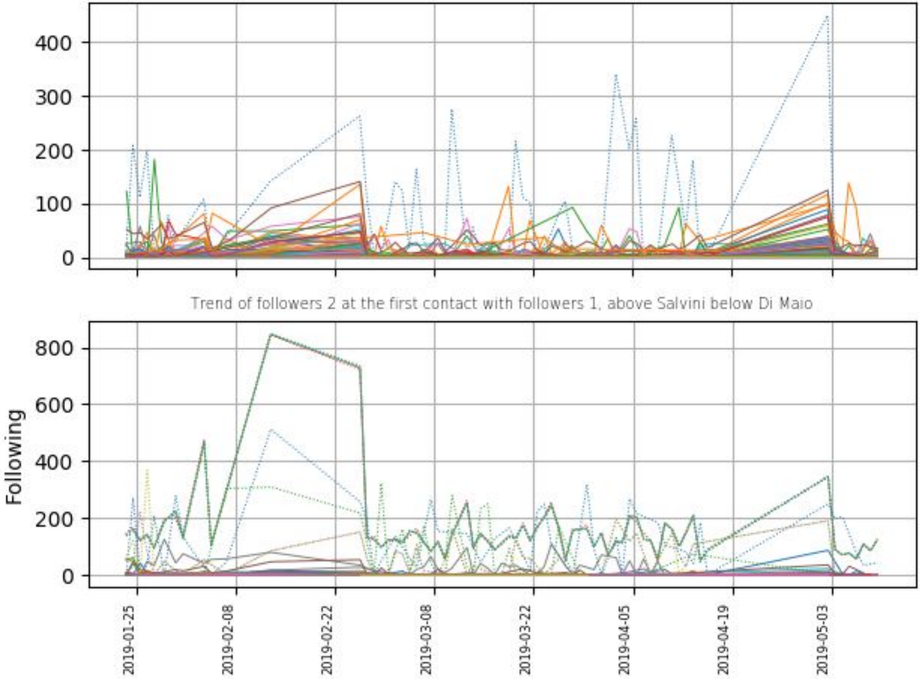


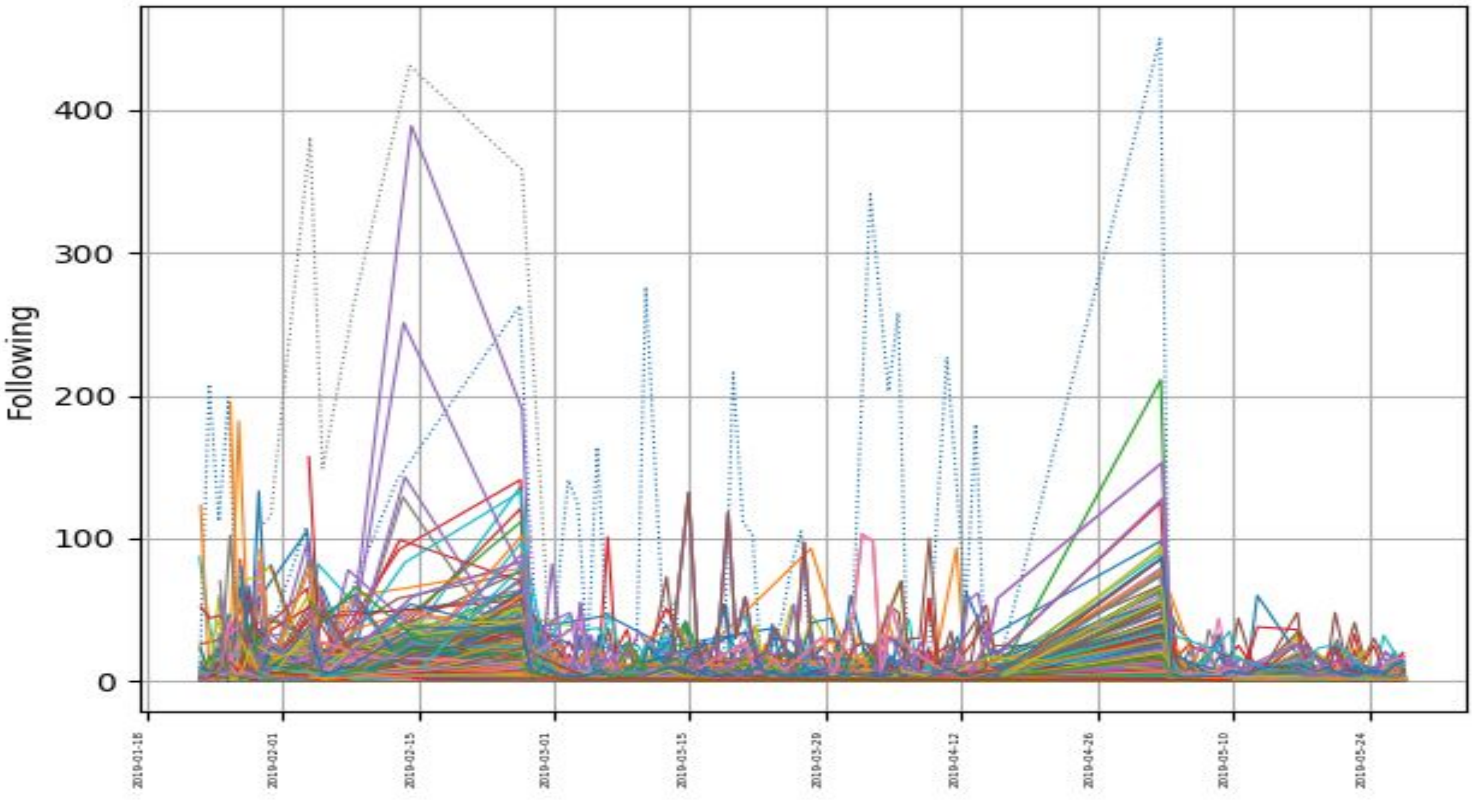
# Our work



# Data analysis

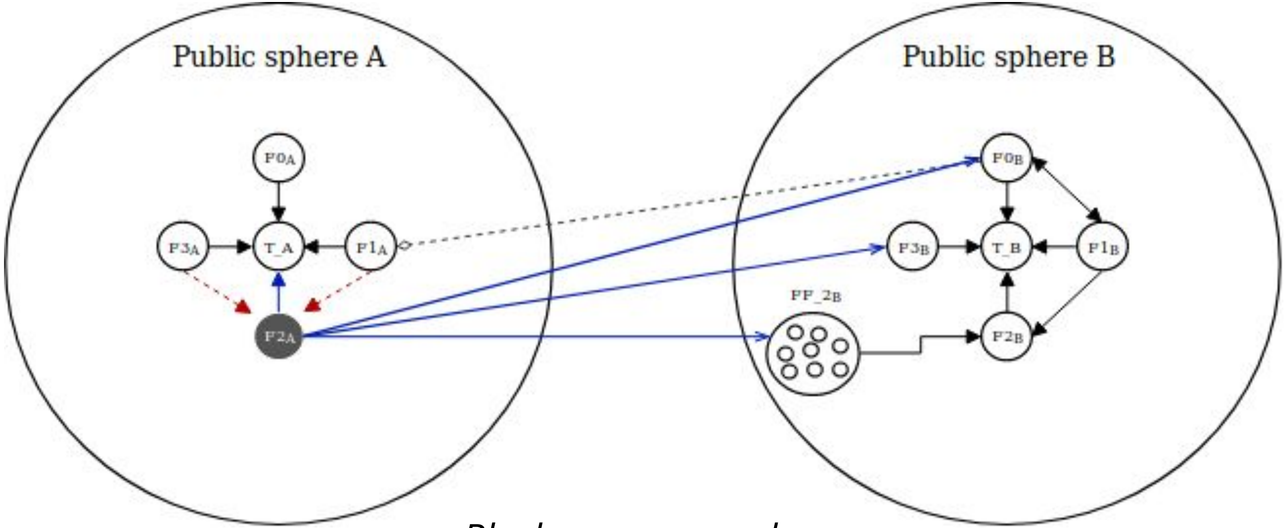



Following
400
300
200
100
0
2019-01-18
2019-02-01
2019-02-15
2019-03-01
2019-03-15
2019-03-29
2019-04-12
2019-04-26
2019-05-10
2019-05-24

# Classes of users

- Targetized
- Bipartisan
- Anomaly_behaviour
- Similar_behaviour
- Most_influential_users
- Black/white users




*Black users example*

# Papers:

Williams, H.T., McMurray, J.R., Kurz, T. and Lambert, F.H., 2015. Network analysis reveals open forums and echo chambers in social media discussions of climate change.
Global Environmental Change, 32, pp.126-138.

Batorski, D. and Grzywińska, I., 2018. Three dimensions of the public sphere on Facebook. Information, Communication & Society, 21(3), pp.356-374.

Garimella, K., Morales, G.D.F., Gionis, A. and Mathioudakis, M., 2018. Political discourse on social media: Echo chambers, gatekeepers, and the Price of bipartisanship. arXiv preprint arXiv:1801.01665.

Leskovec, J., Backstrom, L., Kumar, R. and Tomkins, A., 2008, August. Microscopic evolution of social networks. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 462-470). ACM.

Paranjape, A., Benson, A.R. and Leskovec, J., 2017, February. Motifs in temporal networks. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (pp. 601-610). ACM.

# Papers:

Mark Newman, Albert-Laszlo Barabasi, and Duncan J Watts. The structure and dynamics of networks. Princeton University Press, 2011.

H. Allcott and M. Gentzkow. “Social media and fake news in the 2016 election”. J. Econ. Perspect., 31(2):211–236, May 2017. doi:10.3386/w23089

Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. “Can cascades be predicted?” In Proceedings of the 23rd international conference on World wide web, pages 925–936. ACM, 2014.

Riquelme, F., González-Cantergiani, P. (2016). Measuring user influence on Twitter: A survey. Information Processing & Management. 52, p. 949-975

E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone’s an influencer: quantifying influence on twitter. In Proc. WSDM, 2011

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Fallin Hunzaker, M. B., et al. (2018). Exposure to opposing views on social media can increase political polarization. Proceedings of the National Academy of Sciences, 115(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115.

Wu, Q., Yang, C., Gao, X., He, P. and Chen, G., 2018, November. EPAB: Early Pattern Aware Bayesian Model for Social Content Popularity Prediction. In 2018 IEEE International Conference on Data Mining (ICDM) (pp. 1296-1301). IEEE.

Alfifi, M., Kaghazgaran, P., Caverlee, J. and Morstatter, F., 2019. A Large-Scale Study of ISIS Social Media Strategy: Community Size, Collective Influence, and Behavioral Impact.