The Directed Closure Process in Hybrid Social-Information Networks, with an Analysis of Link Formation on Twitter

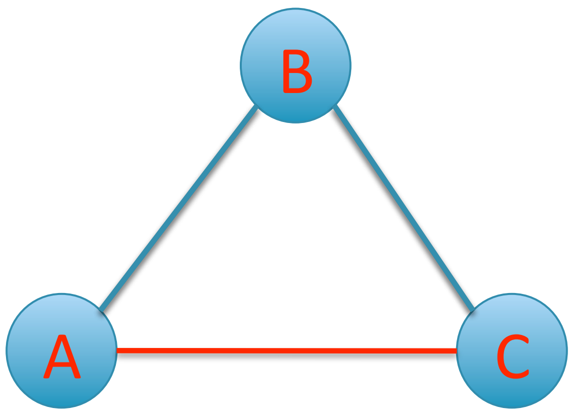
Daniel M. Romero and Jon Kleinberg

Cornell University

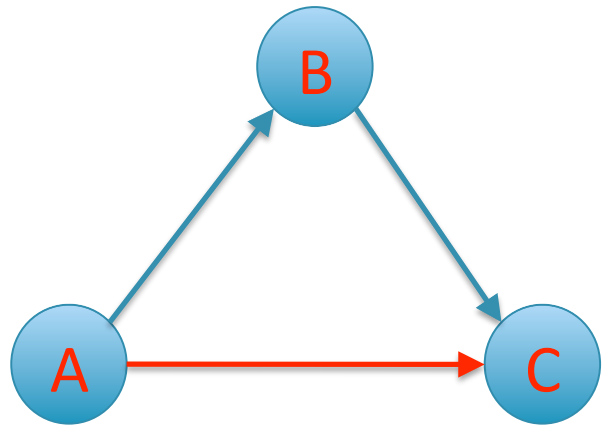
Information vs. Social Networks

- Information Networks: Directed structures, some nodes with extremely large in-degree.
- Social Networks: Roughly undirected, some variation in connectivity but not as much as in Information Networks.
- Twitter: Reflects properties of both.

Triadic Closure vs. Directed Closure



Undirected triangle



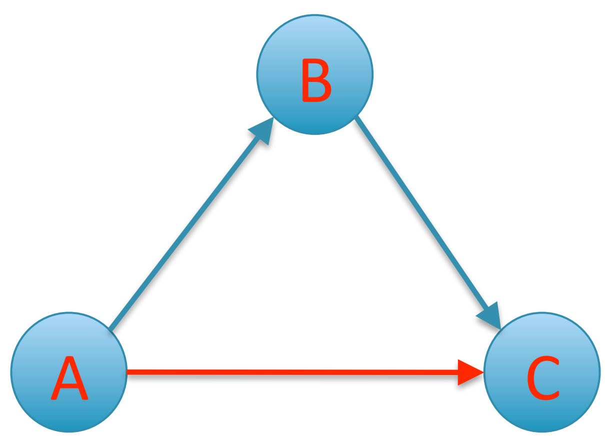
Directed feed-forward triangle

Triadic Closure : An edge connects two nodes who already have a common neighbor

Directed Closure: A node A links to a node C to which it already has a two-step path (through a node B).

Directed Closure

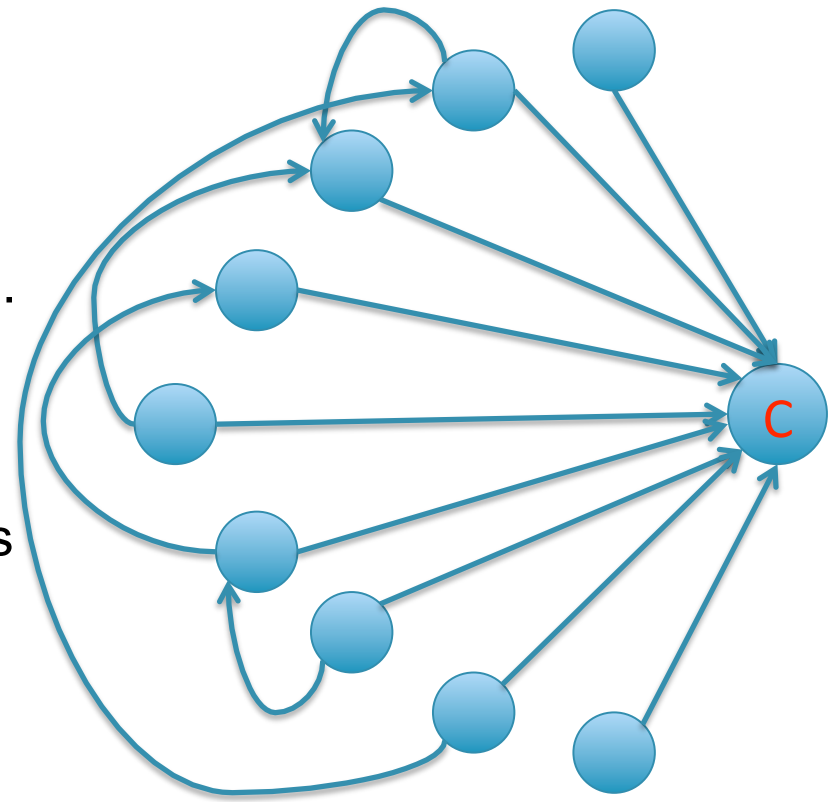
- An edge in a directed graph *exhibits closure* if it completes a two-step path between its endpoints at the time it's formed.



The *closure ratio* node C is the fraction of C's incoming edges that exhibit closure.

- The closure ratio of node C could indicate how many nodes discovered C through other nodes already interested in C.

- The average closure ratio on a network could indicate how much "copying" of edges there is.



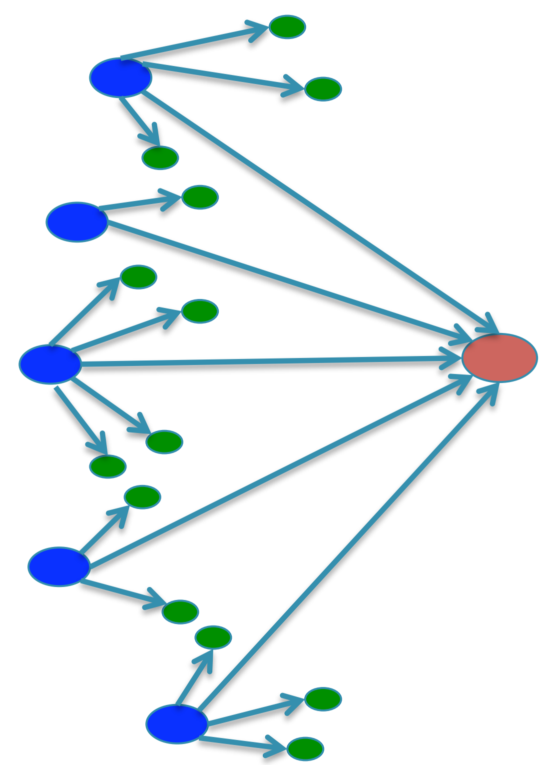
The Data

A random sample of 18 Twitter *micro-celebrities:* Users with between 10,000 and 50,000 followers.

For each node C define:

- $L_{in}(C)$ = Chronologically ordered list of C's followers.

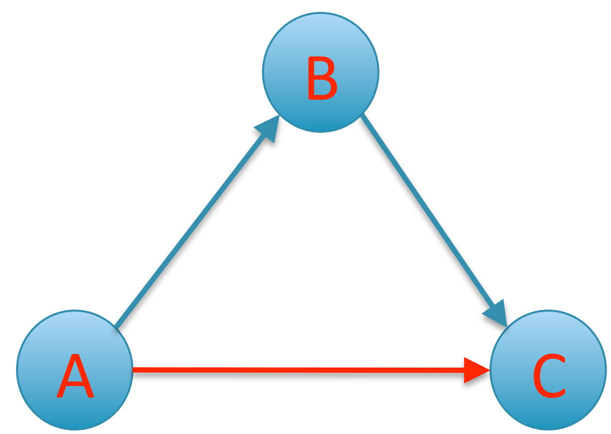
- $L_{out}(C)$ = Chronologically ordered list of the users that C follows.



Can Closure Ratio of Celebrities Be Determined from Data?

(A,C) exhibits closure if:

1. (A,C) was created after (B,C) and

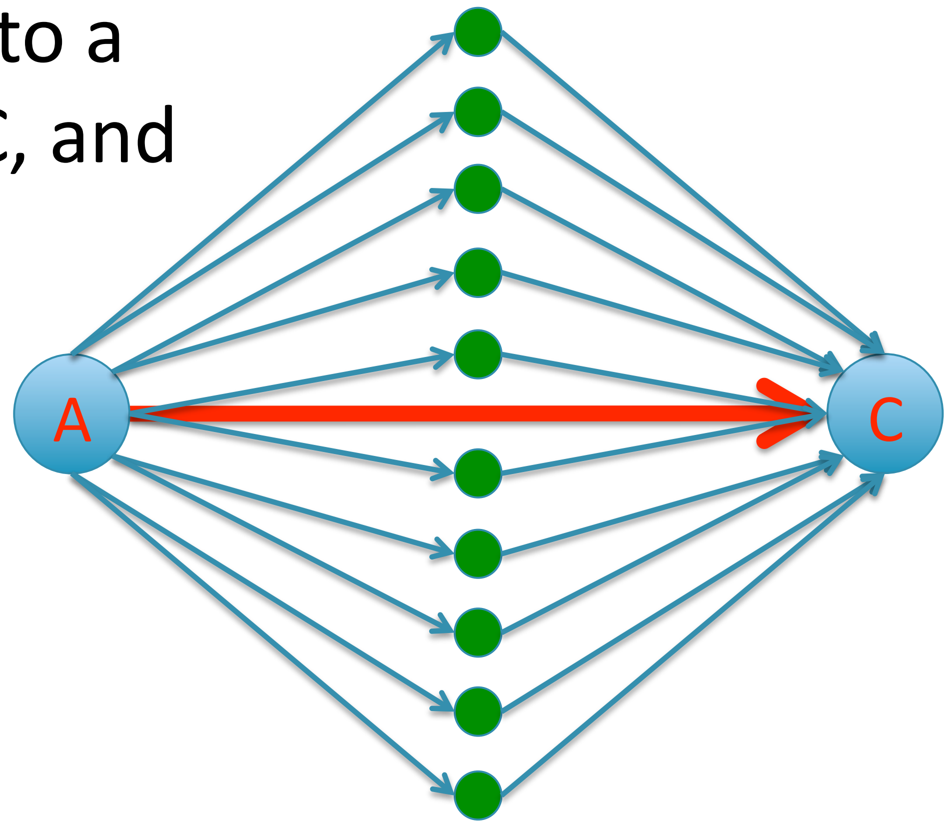


2. (A,C) was created after (A,B).

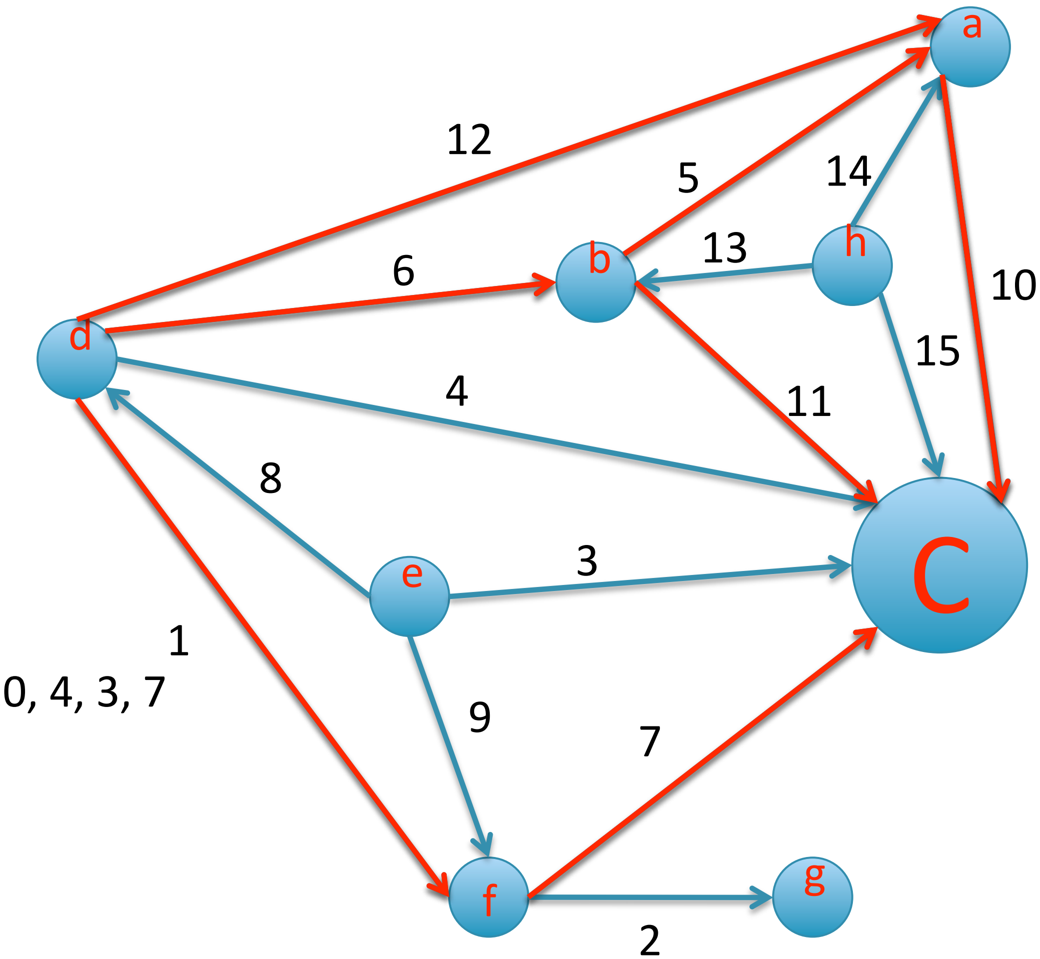
We can determine 1. by looking at $L_{in}(C)$ and 2. by looking at $L_{out}(A)$.

Notation

- A user A is *k-linked* to a user C if A follows C, and A also follows k followers of C.
- Let $S_k(C)$ denote the set of users k-linked to C.
- Let f_k the fraction of users in $S_k(C)$ whose edge to C exhibits closure.



Example



Edges to C that exhibit closure: 15, 11

Edges to C that do not exhibit closure: 10, 4, 3, 7

$S_1(C) = \{b\}, \qquad f_1 = 1$

$S_2(C) = \{h, e\}, \qquad f_2 = \frac{1}{2}$

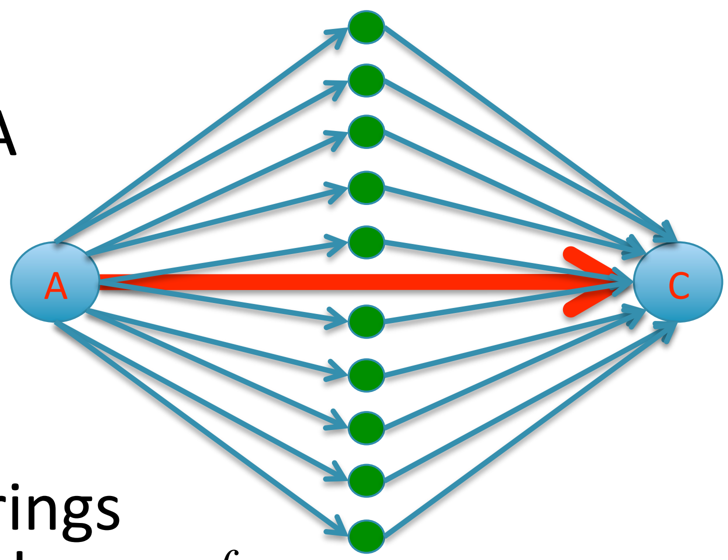
$S_3(C) = \{d\}, \qquad f_3 = 0$

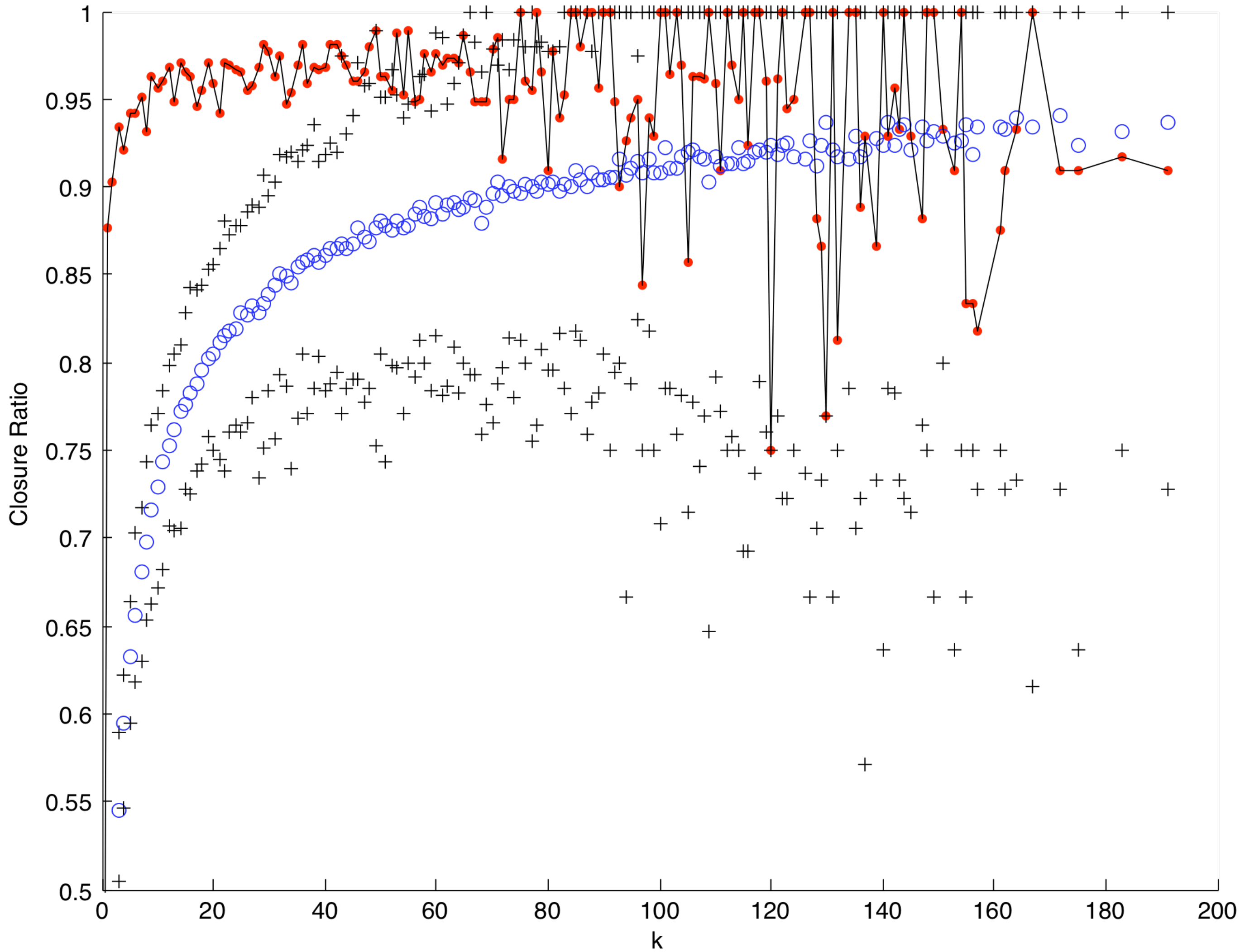
Is Directed Closure a Significant Process?

Randomization Test:

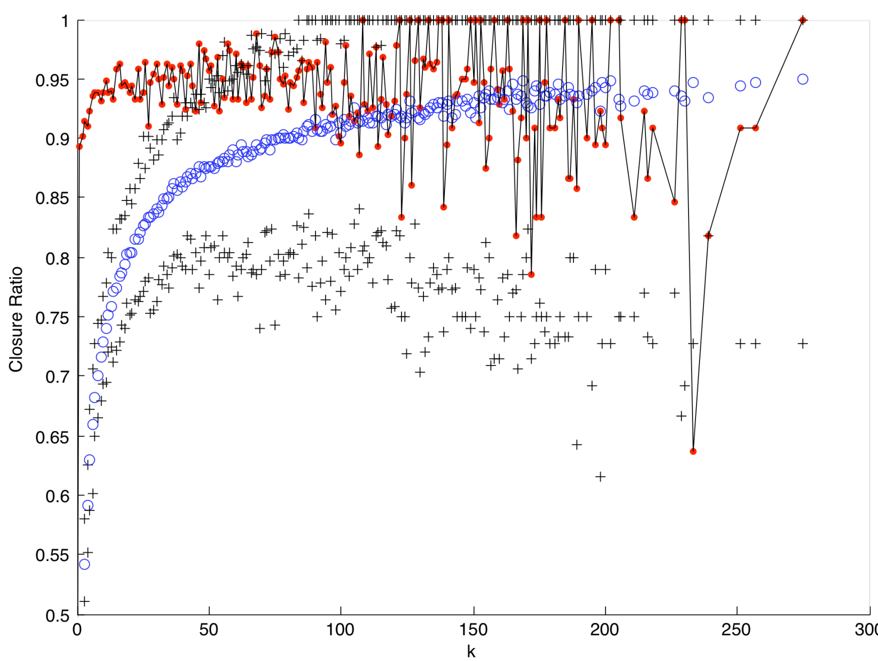
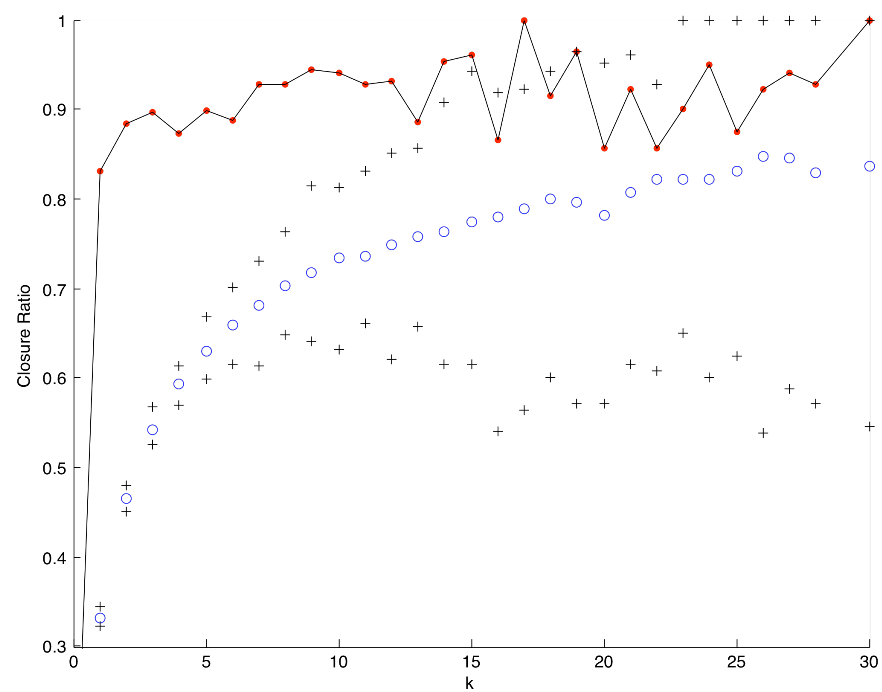
For each celebrity C: For each k with $|S_k|>10$, approximate the expected value of f_k assuming that edges arrive in random order:

1. Generate a network of a node A pointing to a node C and to k other nodes.
2. Choose $|S_k|$ random ordering of the edges of the network.
3. Determine the fraction of orderings in which the edge A-C exhibits closure f_k .
4. Repeat 100 times.



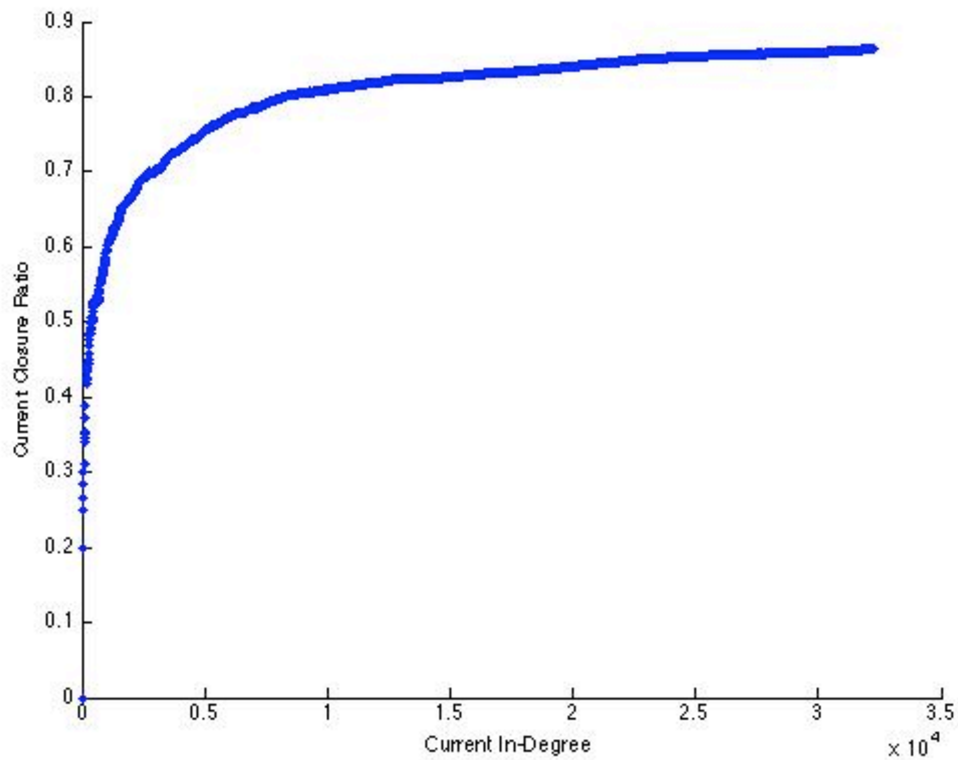


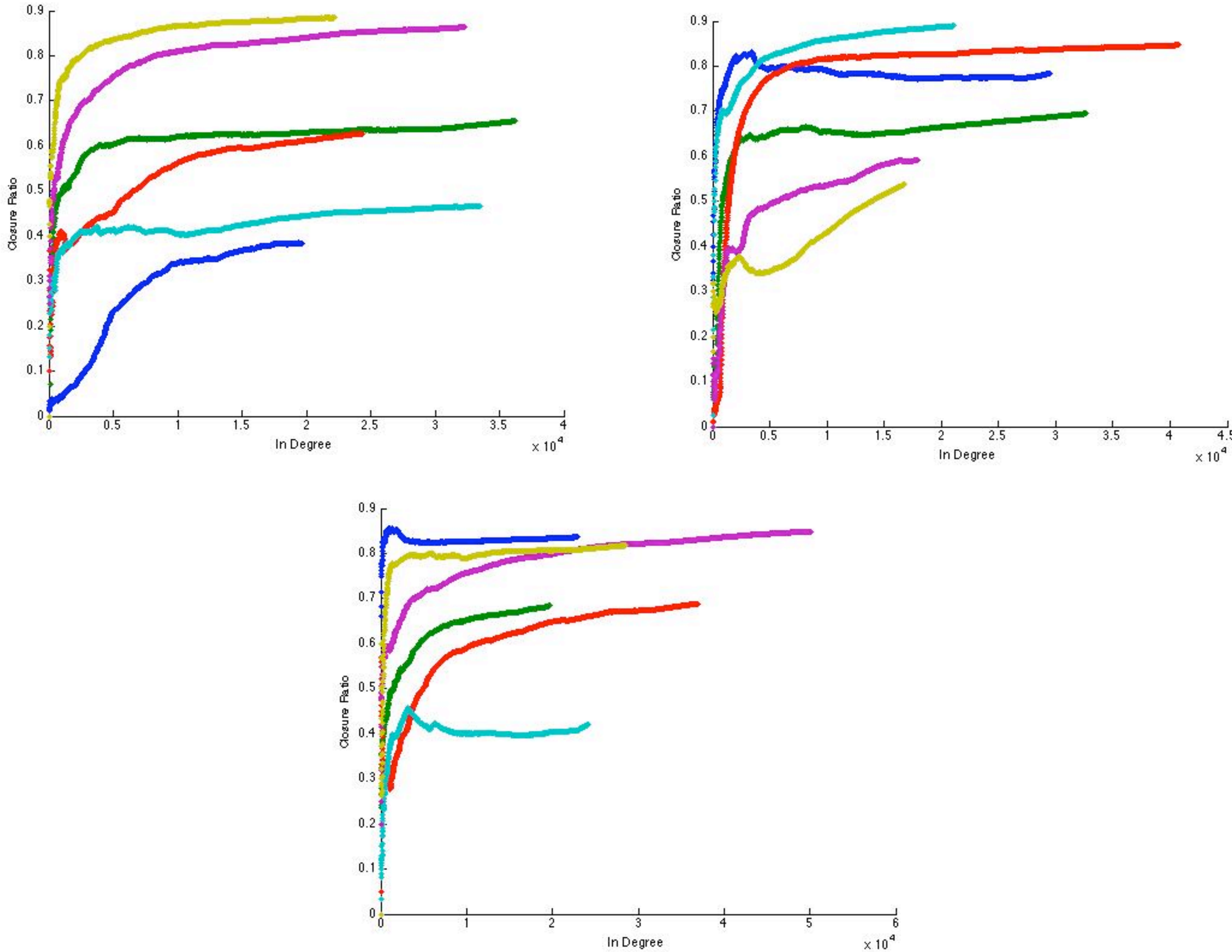
Closure Ratio
k
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0
20
40
60
80
100
120
140
160
180
200


Closure Ratio
k

Closure Ratio
k

Closure Ratio

How does a node's closure ratio change as in degree increases?




0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Closure Ratio
0 0.5 1 1.5 2 2.5 3 3.5 4
In Degree
x 10^4
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Closure Ratio
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5
In Degree
x 10^4
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Closure Ratio
0 1 2 3 4 5 6
In Degree
x 10^4

Properties Observed in Data

For the micro-celebrities studied:

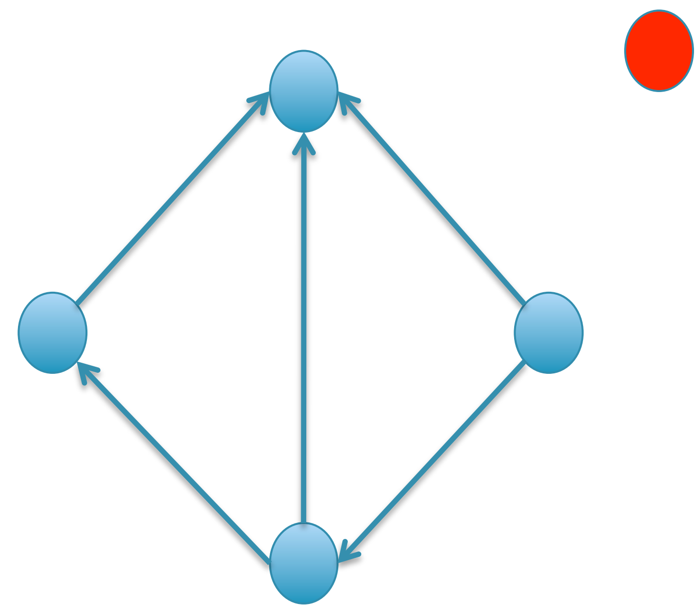
- Closure ratio saturates to a positive constant f
- The constant f is different for different micro-celebrities.
- The constant f is not closely related to the total in-degree of the micro-celebrity.

Can we find a model that captures these properties without a "copying" mechanism?

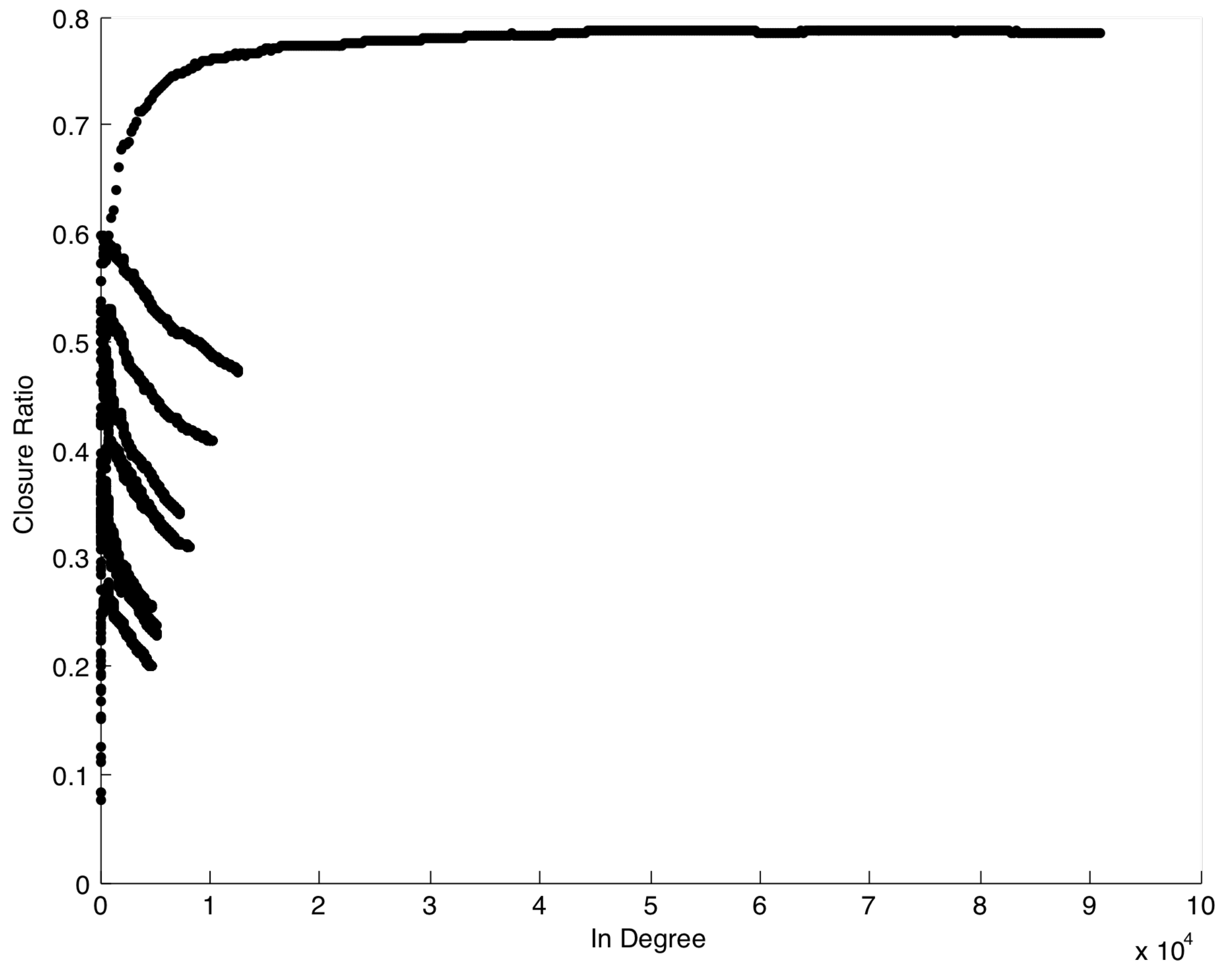
Preferential Attachment

- Fix $\alpha \in [0,1]$, and D, $N \in \mathbb{N}$. The graph will have N nodes labeled 0, 2, ..., N-1.

- Initially ($t = 0$) the graph consists of node labeled 1 with an edge pointing to the node labeled 0.



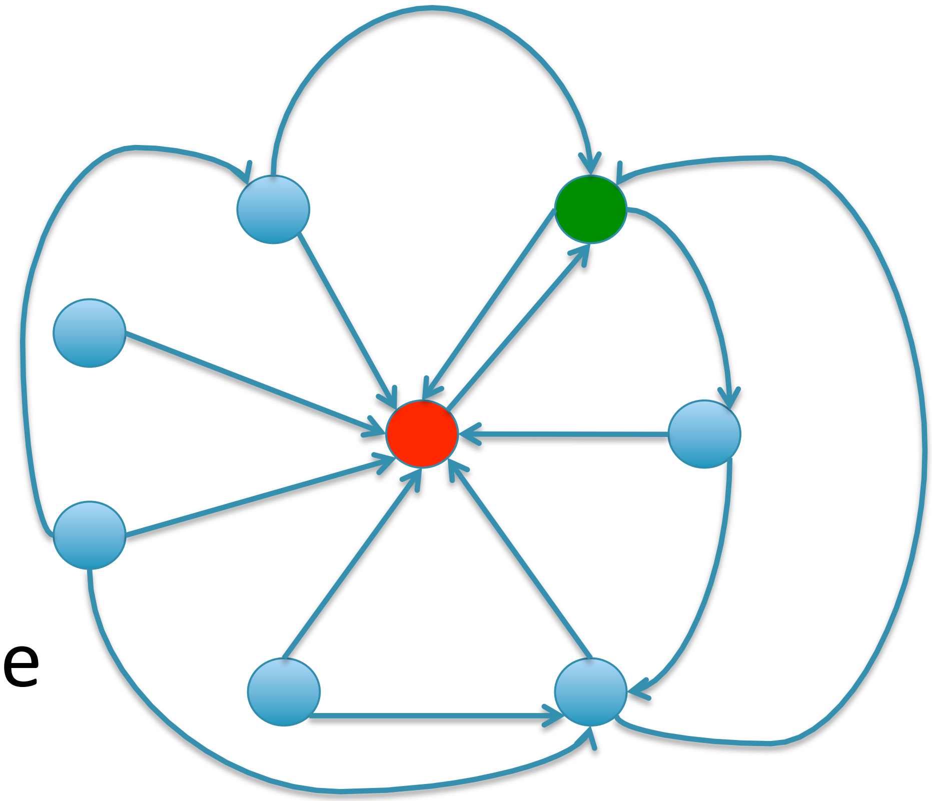
- At each time step ($t = j$) node j will join the graph with D edges directed to nodes chosen from a distribution on 1, 2, ...$j-1$.
  - With probability α choose endpoint uniformly at random
  - With probability $1-\alpha$ weight each node i by d_i



Preferential Attachment with $N = 200{,}000$, $\alpha = .3$, $D = 10$

Heuristic Calculation

- However: A heuristic calculation suggests that the *sum of in-degrees of incoming nodes* plays an important role in determining the node's closure ratio. This is consistent with our data.



Heuristic Calculation

Notation

- $E_t =$ Total Number of edges at time t
- $N_t =$ Total Number of nodes at time t
- $d_t(j) =$ in-degree of node j at time t
- $F_t(j) = \{x : \exists e = (x, j) \text{ at time } t\}$

(Set of nodes that point to j)

- $d_t(S) = \sum_{x \in S} d_t(x)$ (Sum of the

in-degrees of nodes in set S)

- $S_t(j) = \alpha \frac{|F_t(j)|}{N_t} + (1-\alpha)\frac{d_t(F_t(j))}{E_t}$

(Probability that a particular edge from node $t+1$ is directed to a node k which points to j)

$$\boxed{C_{N-1}(j) = 1 - \frac{1-\left(1-S_{N-1}(j)\right)^D}{DS_{N-1}(j)}}$$

- Fix a node j and edge e coming out of node $t+1$.
- Let event $V = \exists\, e' = (t+1, x)$ such that x points to j and e' was created before e.
- Let $P(V) = C_{t,e}(j)$.
- If e is the first edge out of $t+1$ then $C_{t,e}(j) = 0$.
- If e is the second edge out of $t+1$ then $C_{t,e}(j) = S_t(j)$.
- If e is the third edge out of $t+1$ then $C_{t,e}(j) = 1-\left(1-S_t(j)\right)^2$
- If e is the d^{th} edge out of $t+1$ then $C_{t,e}(j) = 1-\left(1-S_t(j)\right)^{d-1}$
- $C_{t,e}(j) = \frac{1}{D}\left[1-\left(1-S_t(j)\right)\right] + \frac{1}{D}\left[1-\left(1-S_t(j)\right)^2\right] + \ldots$

$$+\frac{1}{D}\left[1-\left(1-S_t(j)\right)^{D-1}\right] = 1 - \frac{1-\left(1-S_t(j)\right)^D}{DS_t(j)}$$

- Note that the probability that e exhibits closure is $P(V \mid e = (t+1, j))$
- For the sake of approximation we use $P(V) = C_{t,e}(j)$ as an estimate of the probability that e exhibits closure.
- Thefore if $\lim_{t \to \infty} C_t(j) < \infty$ and N is large enough the final closure ratio of node j is approximately $C_{N-1}(j)$

Improvements to the Model

- Introduce a "fitness" parameter that represents a node's attractiveness.
- Break up nodes into communities and assume nodes are more likely to attach to nodes from their own community.
- The variable *sum of in-degrees of incoming nodes from the same community* is important in determining closure ratio.

Conclusion

- Definition and methodology for directed closure
- Evidence for directed closure in Twitter
- Found explanation for findings through preferential attachment models
- Identified subtle parameter related to closure ratio – the sum of the in-deg of one's followers

Further Directions

- Identify other causes of significant number of edges exhibiting closure.
- Identify communities on Twitter in order to test predictions of the preferential attachment with communities model.
- Find the analogous of directed closure in other social and information networks and compare measures.