

# Predicting Influentials in Online Social Networks

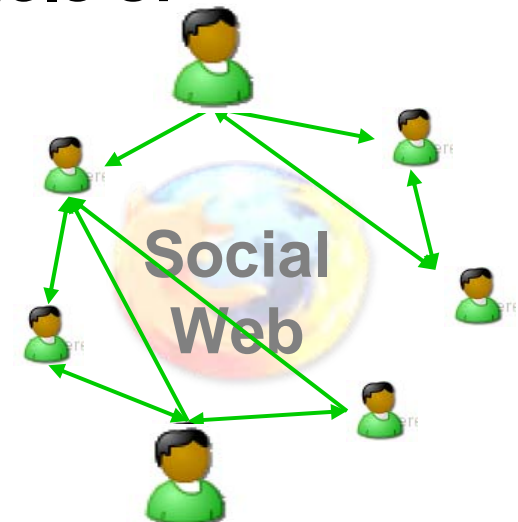
**Rumi Ghosh**

**Kristina Lerman**

**USC Information Sciences Institute**



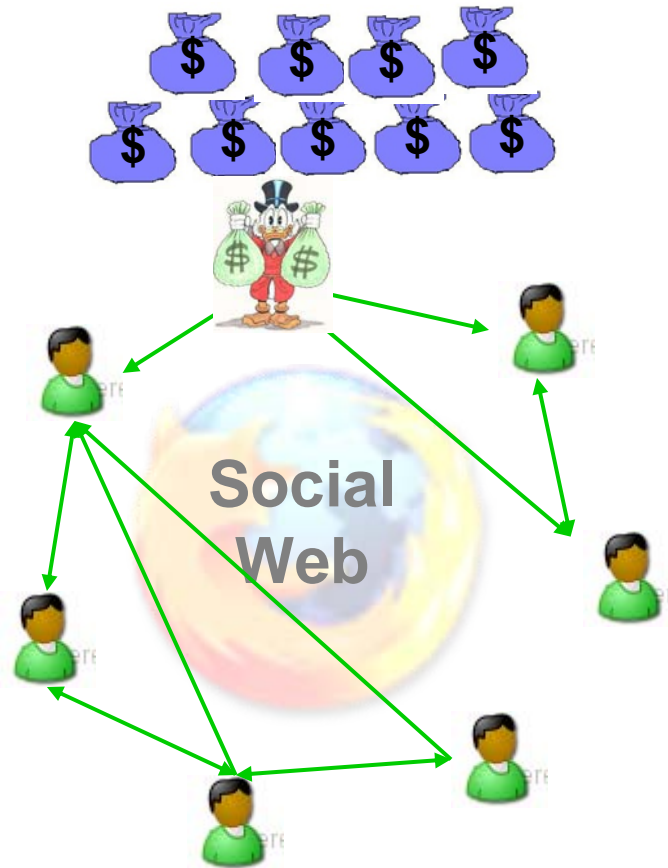
- **WHO is IMPORTANT?**
  - Characteristics
    - *Topology*
    - *Dynamic Processes /Nature of flow*
- **What are the suitable METRICS that can be used to PREDICT influentials in social network?**
  - Characteristics
    - *Dynamic Process*
- **How do we EVALUATE predictive models of influence?**



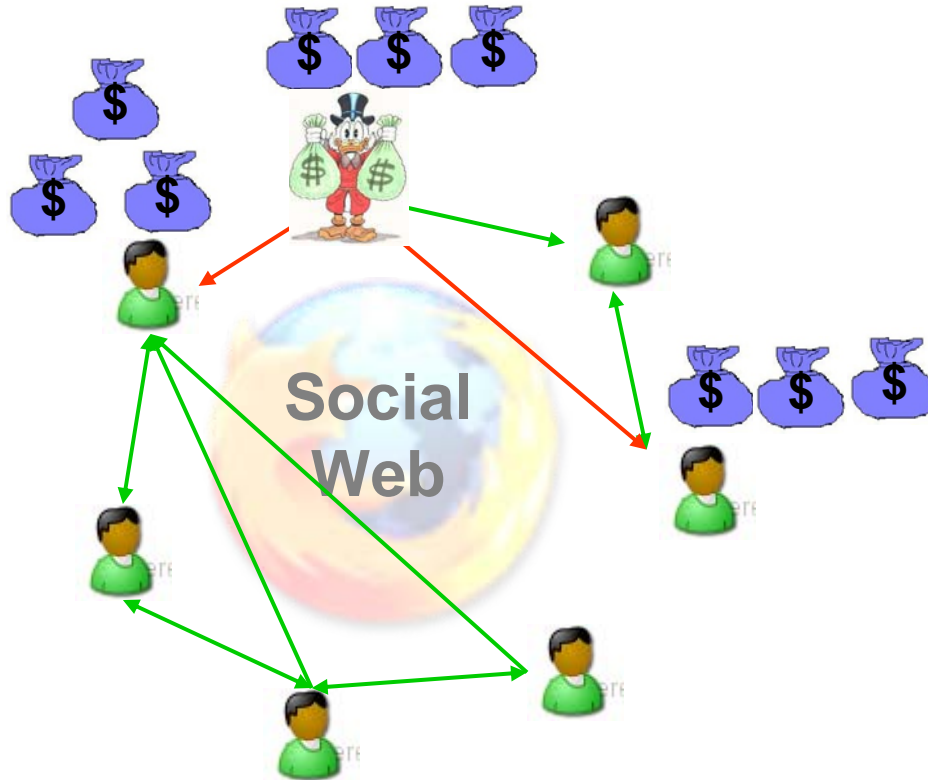
- **Prediction of Influence**
  - Classification of influence models :
    - *Conservative and Non-conservative*
    - *The details of the underlying dynamic process on a network should match those of the influence model.*
      - *Page Rank is not always the best!*
- **Evaluation of Influence**
  - Empirical Measure of Influence (Statistically Significant)
    - *Social News Aggregator Digg*
    - *Dynamic Process-Information propagation*
  - First work evaluating predictive models of influence, using the actual dynamic process, occurring in a social network
- **Mathematical formulation and analytical proofs**
  - Normalized  $\alpha$ -Centrality

- **Classification of Dynamic Processes on Networks**
  - Conservative
  - Non-conservative

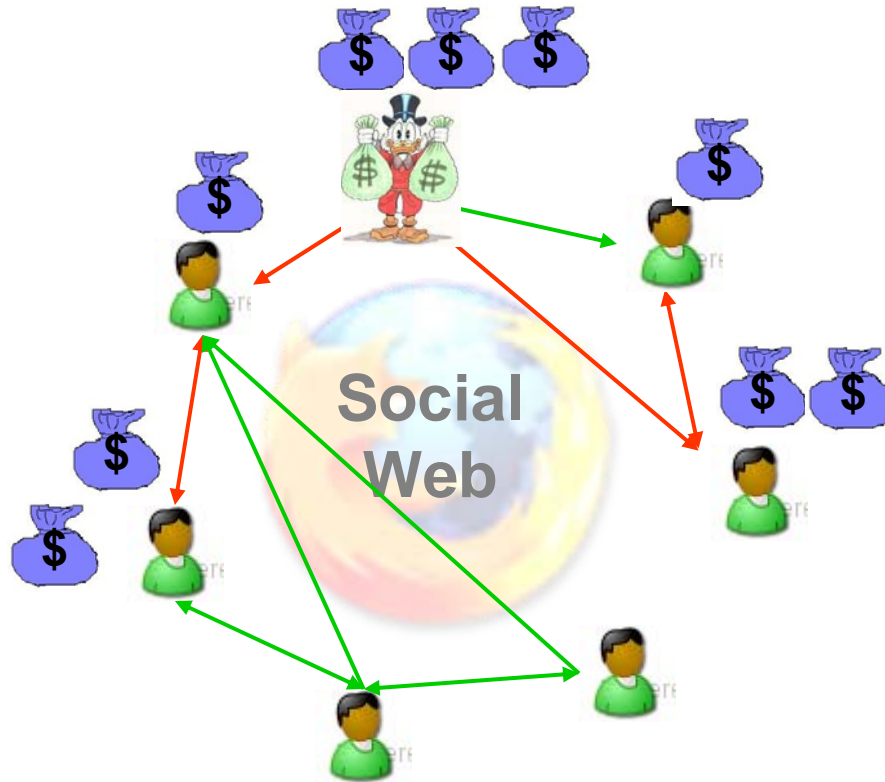
Blue Bags in the network=8



Blue Bags in the network=8

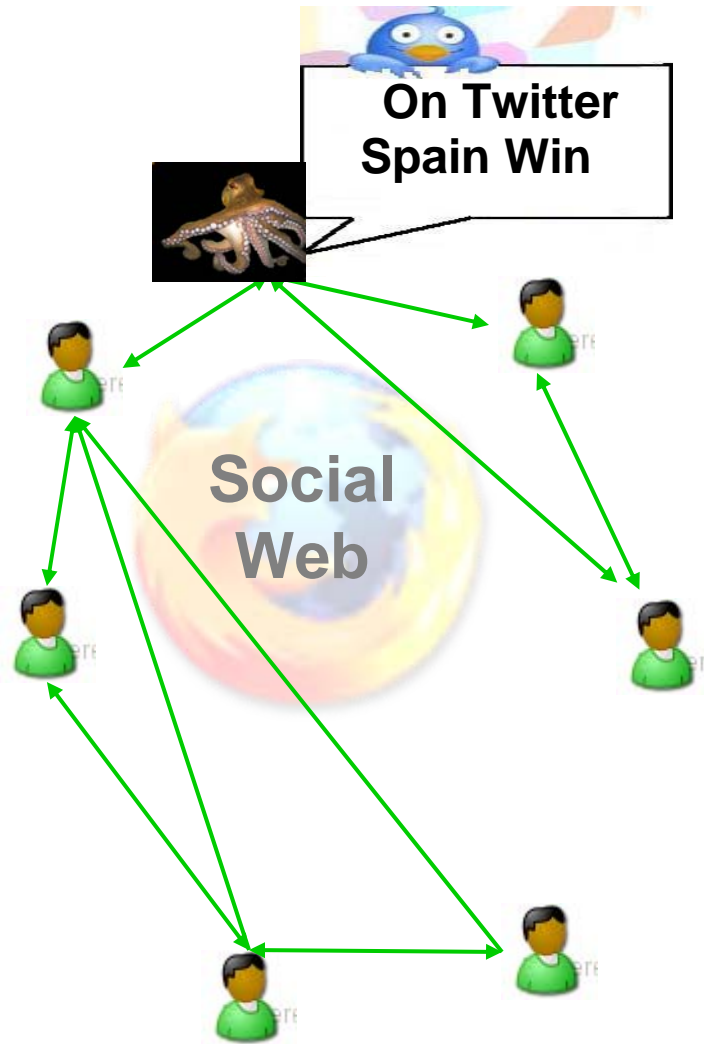


Blue Bags in the network=8

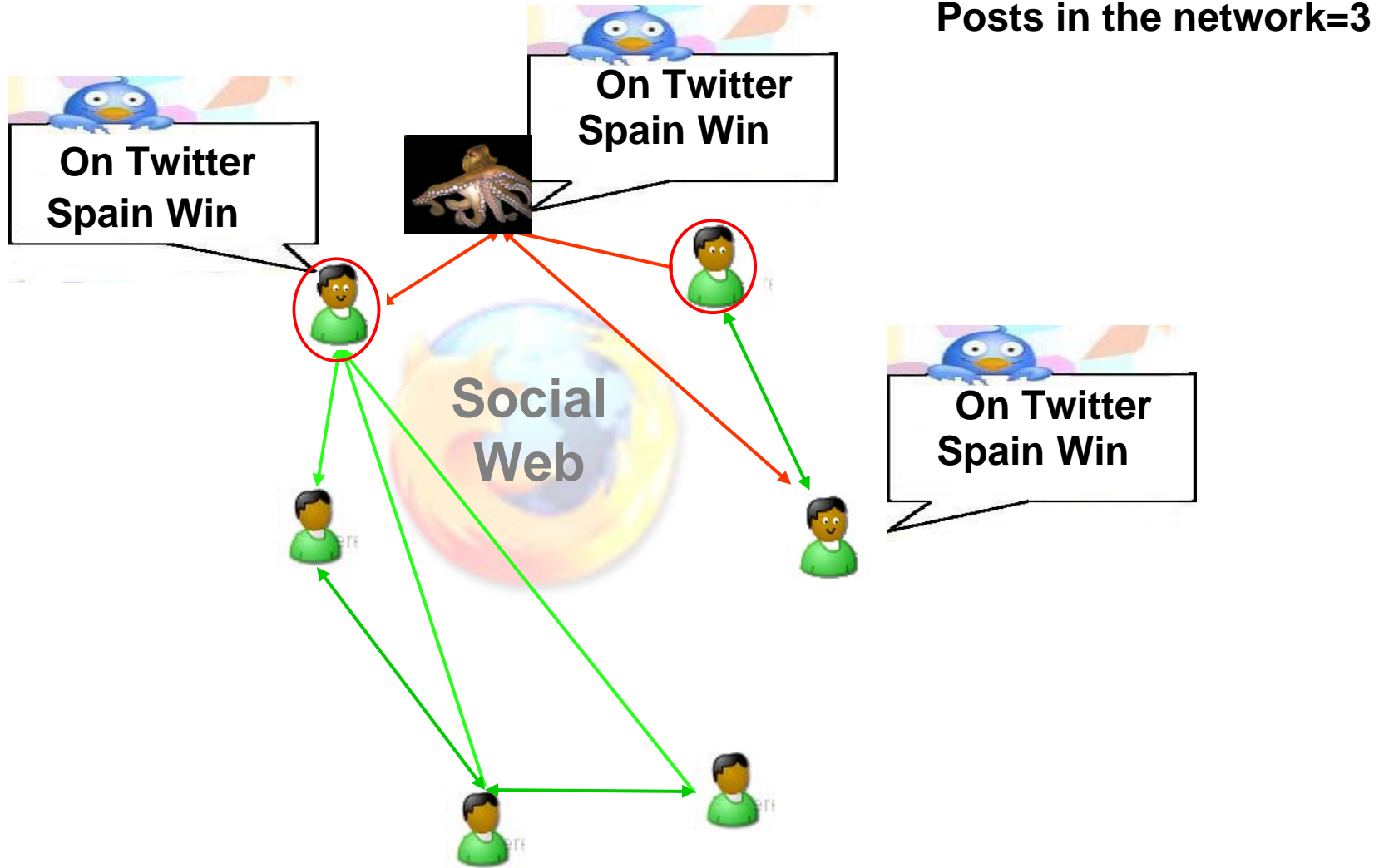


# Non-Conservative Process

Posts in the network=1

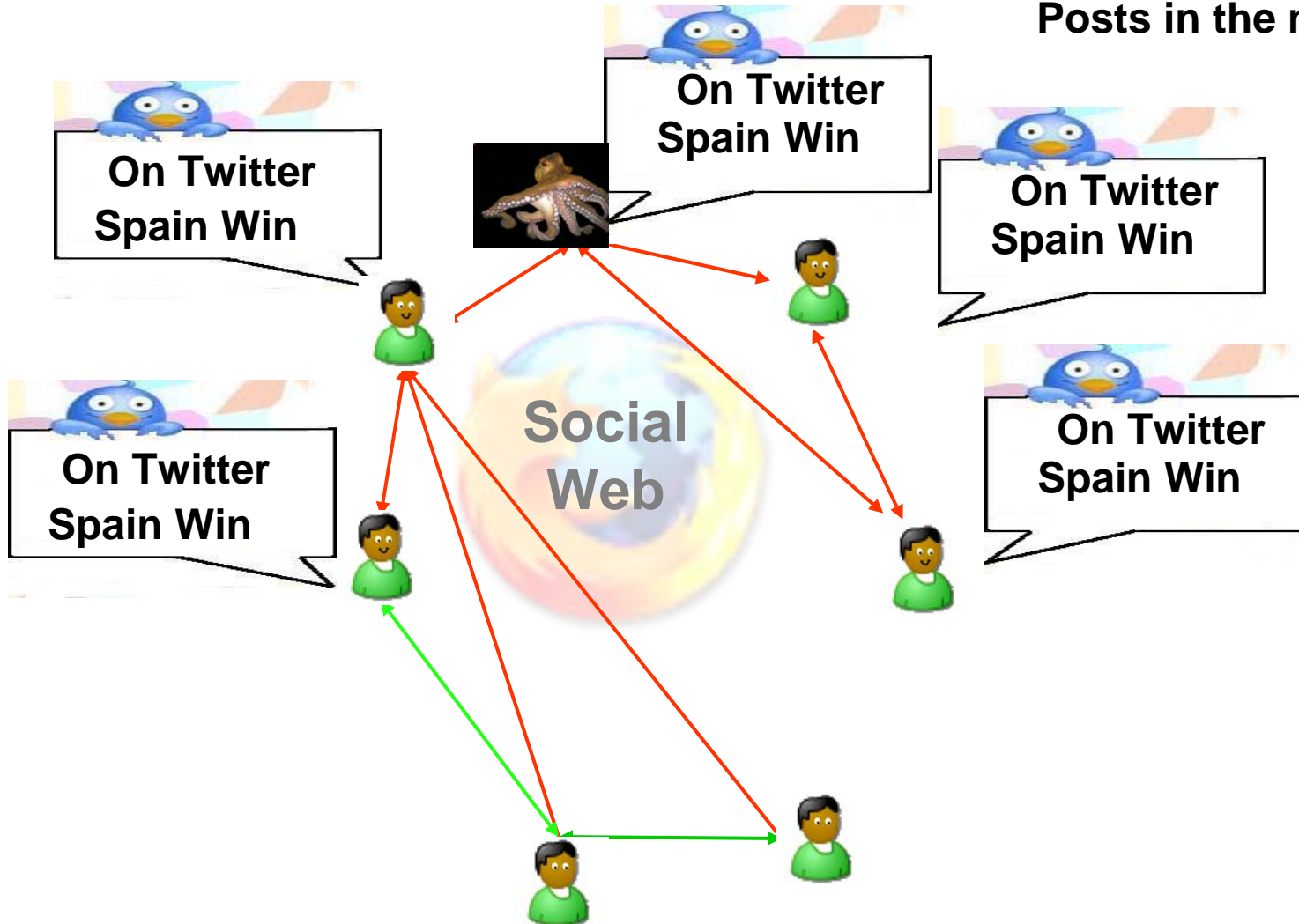


# Non-Conservative Process



# Non-Conservative Process

Posts in the network=5



- Exists empirical studies, structural models
- Two ways to quantify influence
  - 1 Empirically measure online social behavior or dynamic processes to *estimate* influence [Lee, Cha]
  - 2 Use influence models (centrality metrics) based on the structure of the underlying social network to *predict* influence.
- We evaluate predictive influence models using empirical measures of influence.

- Geodesic Path Based ranking measures
  - *Closeness centrality* [Hakimi, Sabidussi, Wassermann et al, Lin]
  - *Graph centrality* [Hage et al.]
  - *Betweenness centrality* [Freeman]
- Topological ranking measures
  - *Markov Process Based Ranking Measures*
    - *Page Rank* [Brin et al.]
    - *Hubbel's Model*
  - *Degree Centrality*
    - *In-degree*
    - *Out-degree Centrality*
  - *Path-Based Ranking Measures*
    - *$\alpha$ -centrality* [Bonacich]
    - *normalized  $\alpha$ -centrality*
    - *Katz score* [Katz]
    - *SenderRank* [Kiss et al.]
    - *EigenVector centrality* [Bonacich]

## Non-Conservative

### Topological Ranking Measures

- *Degree Centrality*
  - *In-degree Centrality*
  - *Out-degree Centrality*
- *Path-Based Ranking Measures*
  - *$\alpha$ -centrality*
  - *Normalized  $\alpha$ -centrality*
  - *Katz Score*
  - *SenderRank*
  - *Eigenvector Centrality*

## Conservative

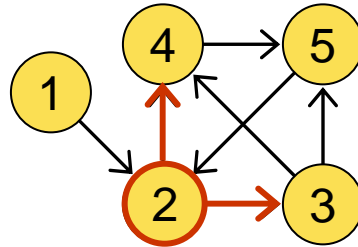
### Geodesic Path-Based Ranking Measures

- *Closeness Centrality*
- *Graph Centrality*
- *Betweenness Centrality*

### Topological Ranking Measures

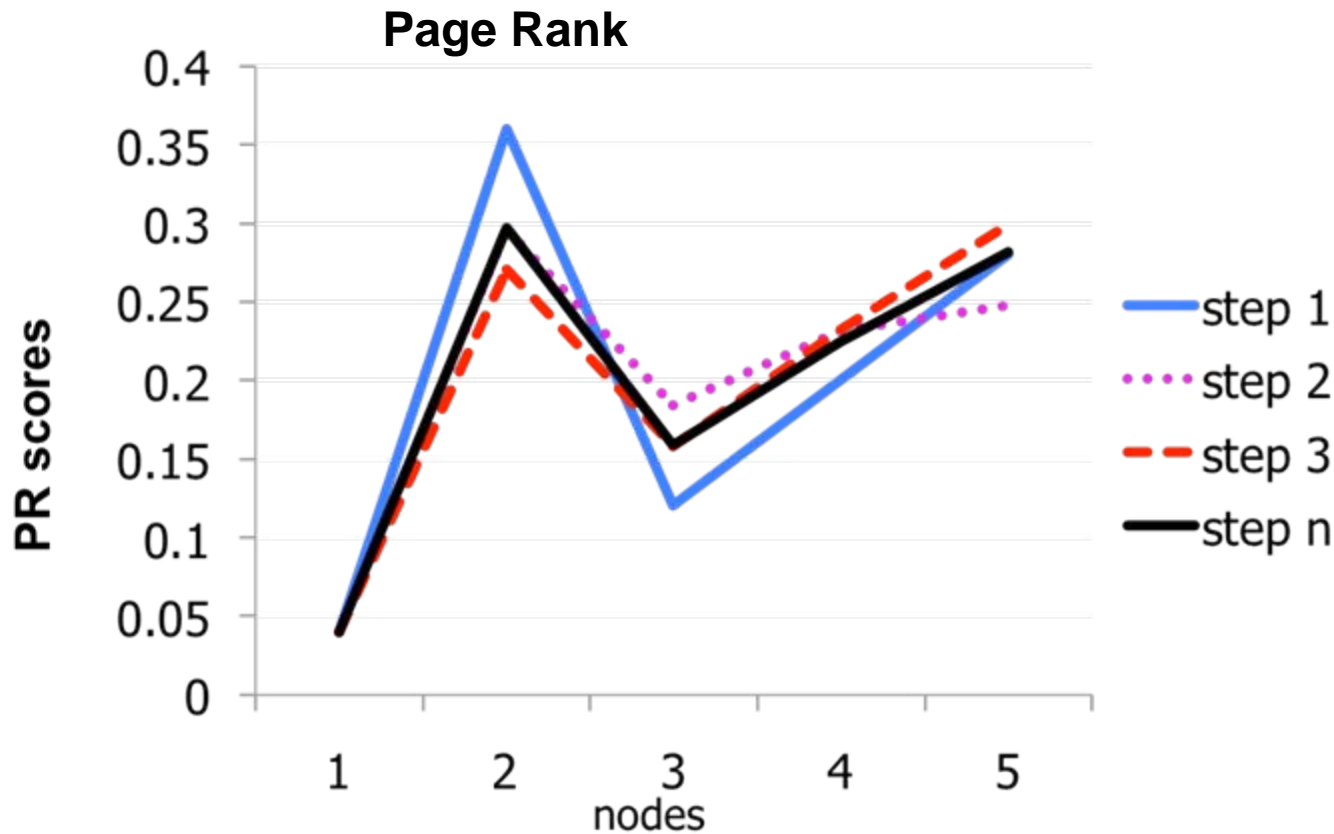
- *Markov Process Based Ranking Measures*
  - *Page Rank*
  - *Hubbel's Model*

# Page Rank

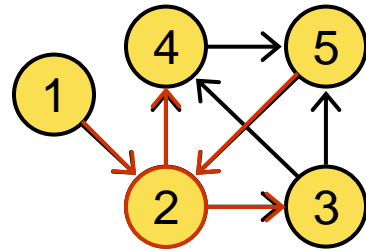


$$C_{pr,\alpha}(i) = (1-\alpha)\frac{1}{n} + \alpha \sum_{j \in \text{fan}(i)} \frac{C_{pr,\alpha}(j)}{d_j^{\text{out}}}$$

[Brin]



# Degree Centrality

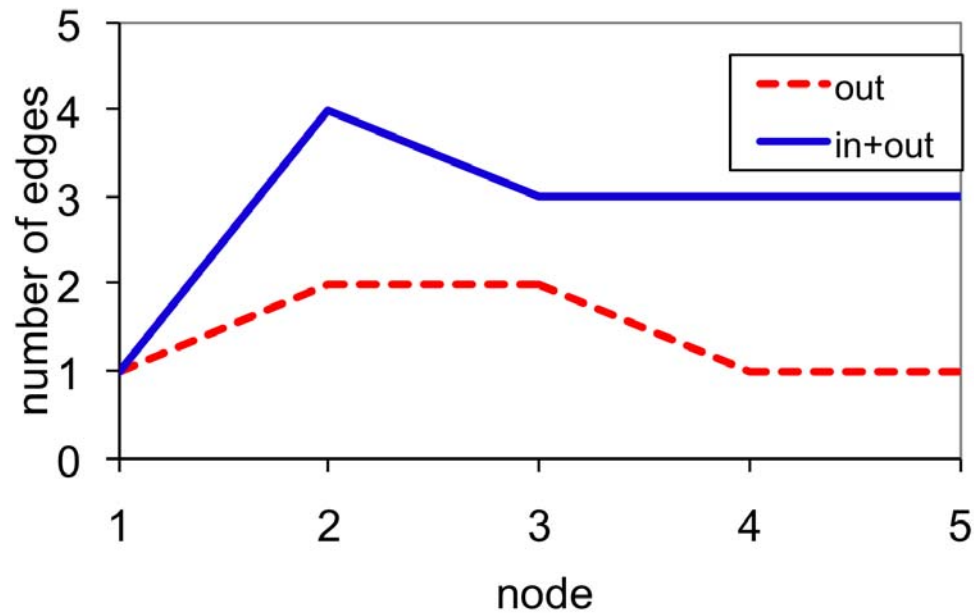


$$C_{d^{in}}(i) = d^{in}(i)$$

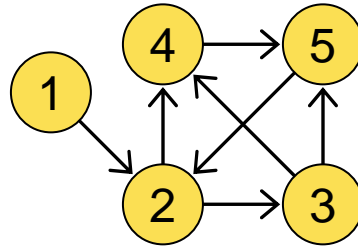
$$C_{d^{out}}(i) = d^{out}(i)$$

$$C_{d^{in+out}}(i) = d^{in}(i) + d^{out}(i)$$

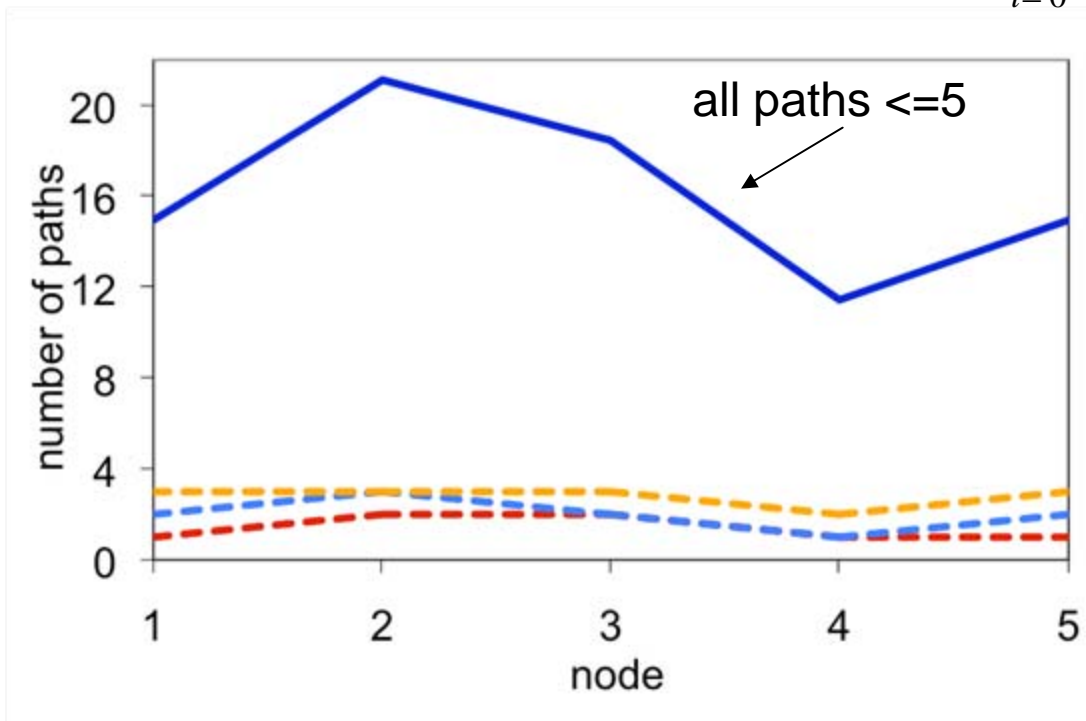
Degree centrality



# $\alpha$ -centrality (Bonacich)



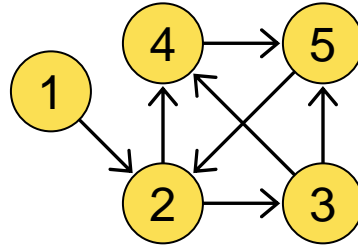
$$C_{\alpha, k \rightarrow \infty} = A + \alpha A^2 + \dots + \alpha^n A^{n+1} + \dots = \sum_{i=0}^{k \rightarrow \infty} \alpha^i A^i \text{ where } \alpha < \frac{1}{|\lambda_1|}$$



Parameter  $\alpha$  sets the length scale of interactions

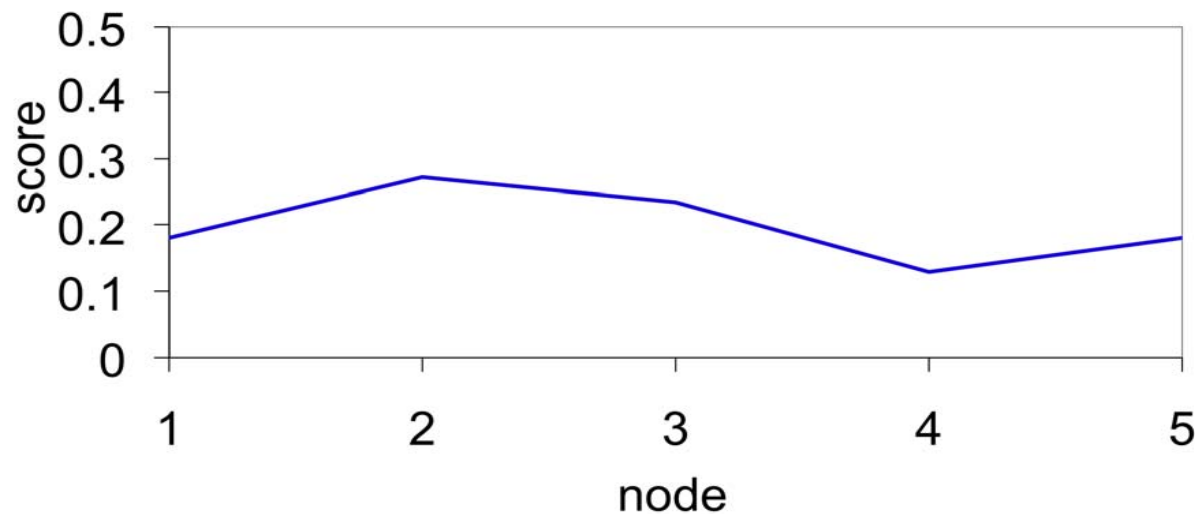
Mean path length =  $(1-\alpha)^{-1}$

# Normalized $\alpha$ -centrality



$$C_{\alpha, k \rightarrow \infty} = \sum_{i=0}^{k \rightarrow \infty} \alpha^i A^i \quad NC_{\alpha, k \rightarrow \infty} = \frac{1}{\sum_{i,j} (C_{\alpha, k \rightarrow \infty})_{i,j}} C_{\alpha, k \rightarrow \infty}$$

Normalized centrality score



$\alpha$ -centrality Matrix:

$$C_{\alpha,k} = \sum_{t=0}^k \alpha^t A^t$$

$\alpha$ -centrality:

$$C_{\alpha} = \nu C_{\alpha,k \rightarrow \infty} = \nu (I - \alpha A)^{-1}$$

where  $\alpha < \frac{1}{|\lambda_1|}$

**Normalized  $\alpha$ -centrality:**

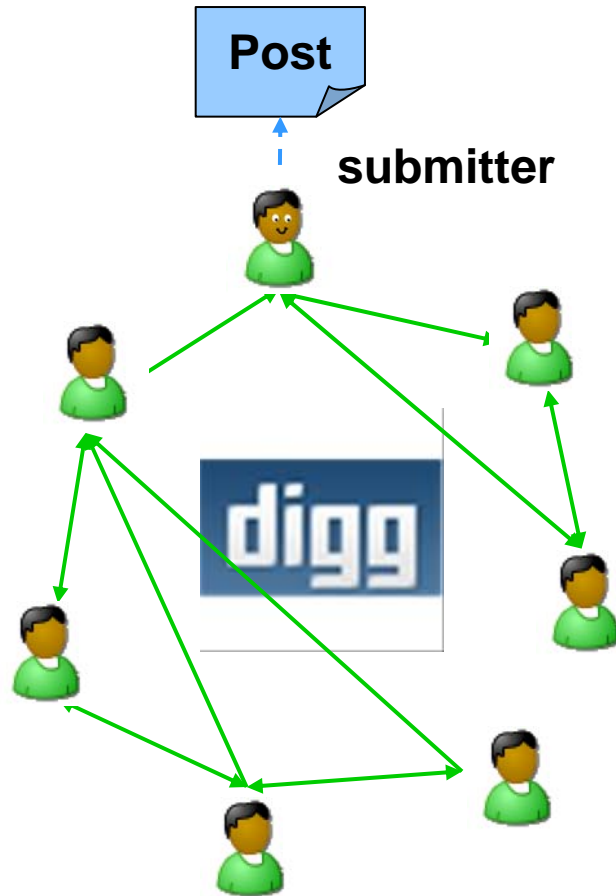
$$NC_{\alpha,k \rightarrow \infty} = \frac{\nu C_{\alpha,k \rightarrow \infty}}{\sum_{i,j} (C_{\alpha,k \rightarrow \infty})_{ij}}$$

- Simple Algorithm
- Does not depend on eigenvalue computation (unlike  $\alpha$ -centrality)
- We give analytical proofs and conditions for
  - Equivalence of ranking due to **normalized  $\alpha$ -centrality** and  **$\alpha$ -centrality**
  - Equivalence of ranking due to **eigenvector centrality** and **normalized  $\alpha$ -centrality**
  - Convergence of normalized  $\alpha$ -centrality
  - Criteria for parametric independence of normalized  $\alpha$ -centrality
  - Other analytical proofs for limiting conditions and boundary values.

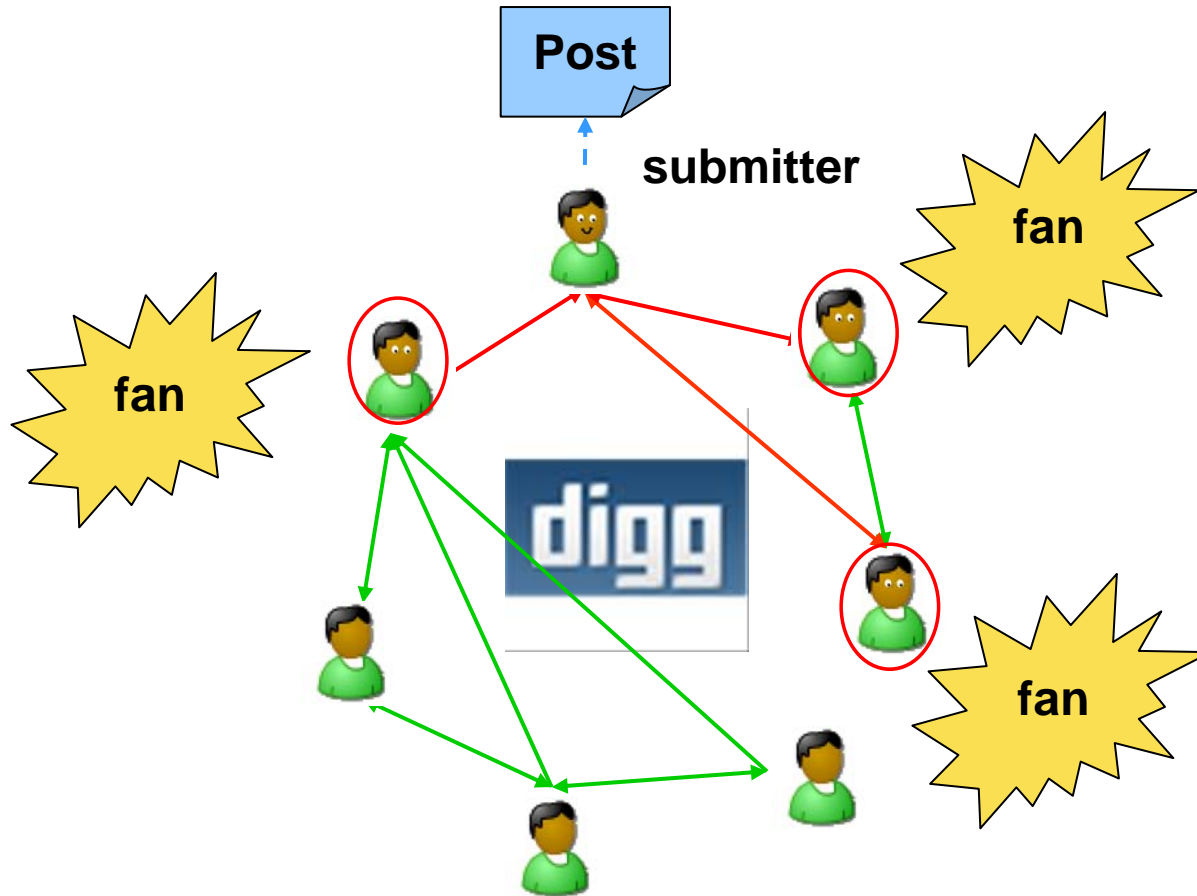


Which model best predicts  
influentials?  
Evaluation?

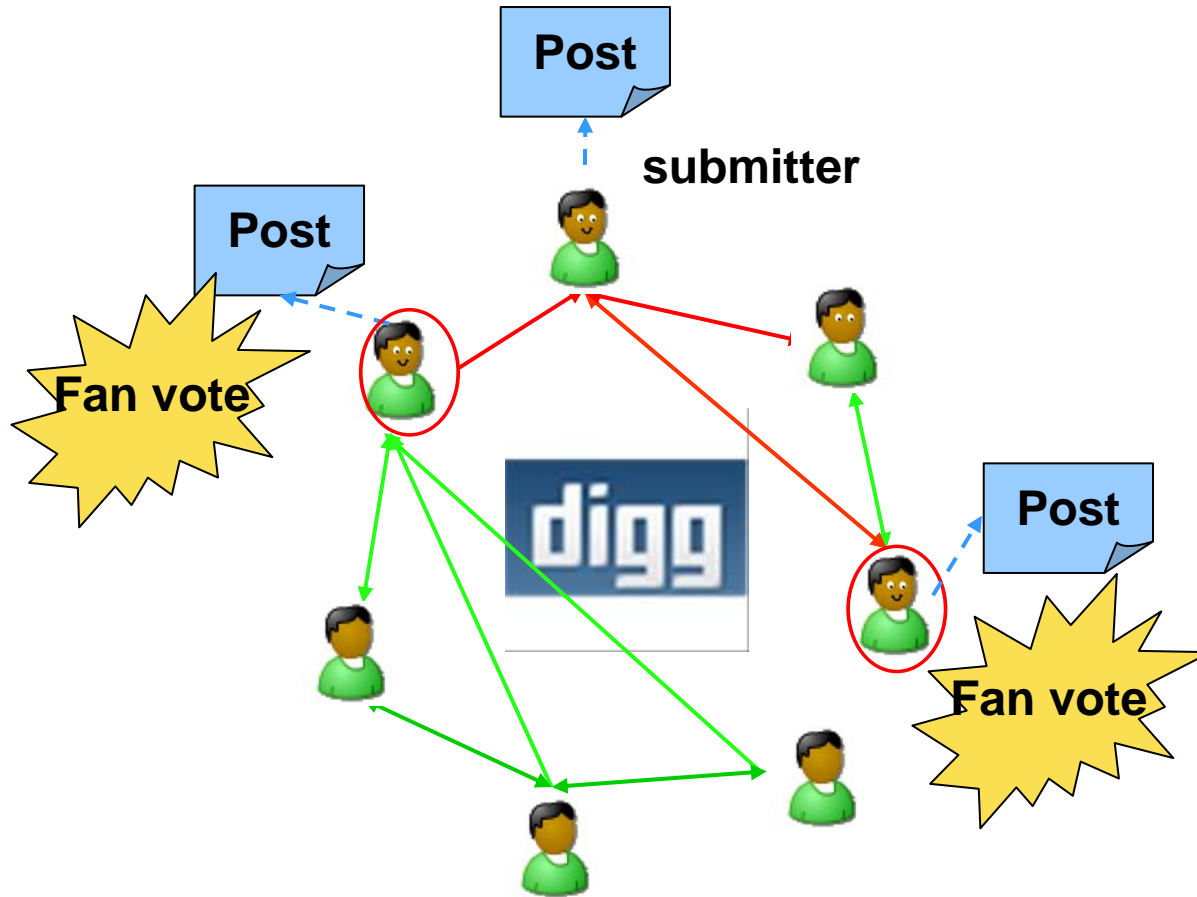
# Information Flow on Digg

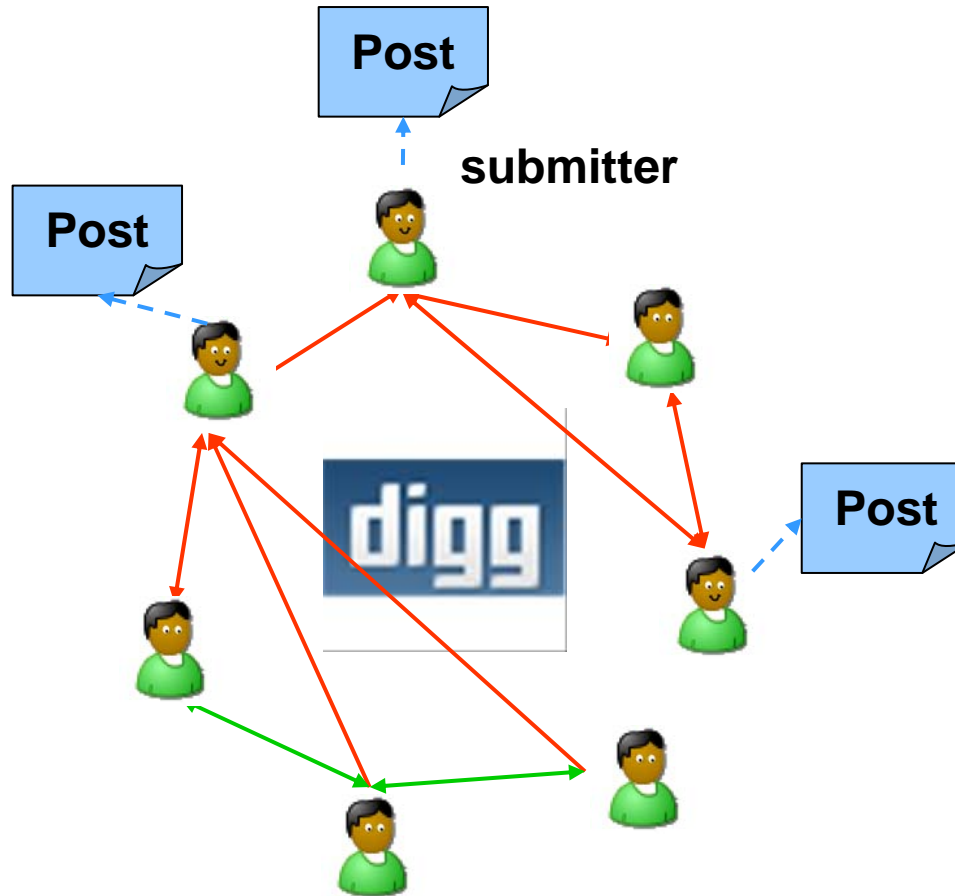


# Information Flow on Digg



# Information Flow on Digg

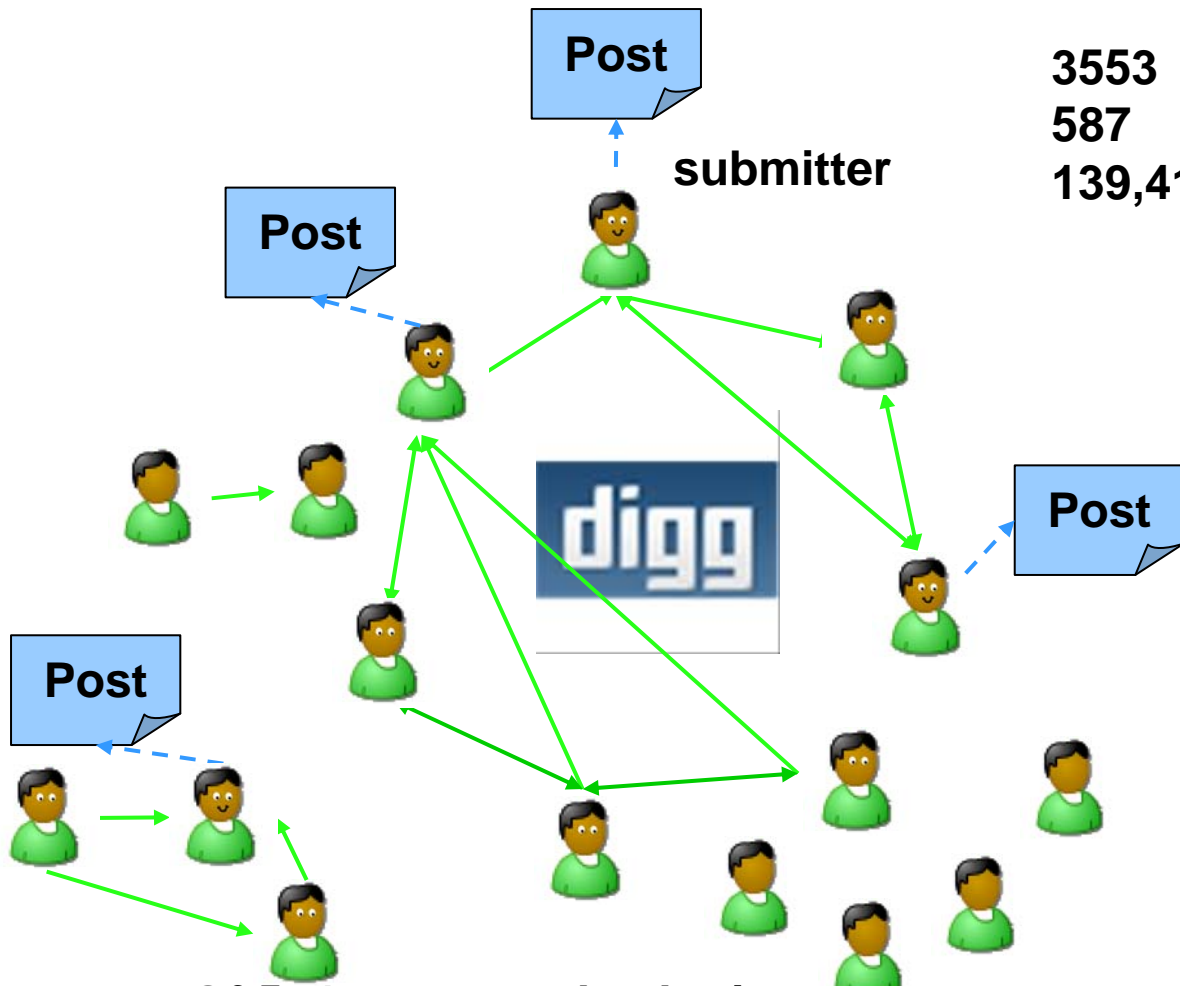




## Non-Conservative Information Propagation on Digg

Hypothesis: non-conservative influence model best predicts influentials

# Data Collection-Digg



**3553 stories**  
**587 distinct submitters**  
**139,410 distinct voters**

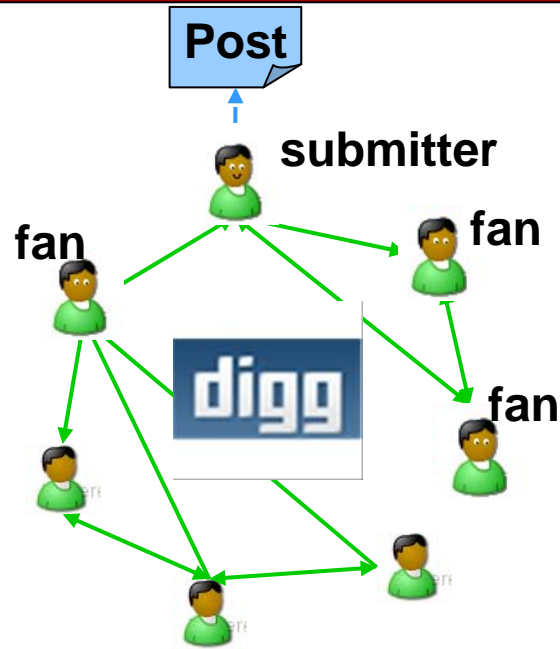
**Social network:**  
**voter connected to**  
**at one or more**  
**voters**  
**69,524 connected**  
**voters**

**Of 574 connected submitters**  
**belonging to the friendship**  
**network, 504 submitters received**  
**at least 1 fan vote in first 100 votes**  
**(in at least 1 story).**

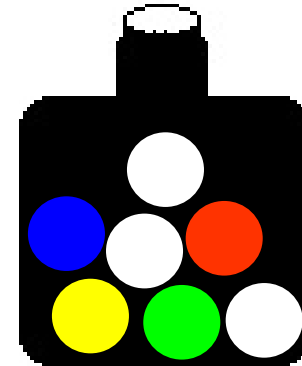
**574 connected submitters**  
**3489 stories**

- **Probability of a fan vote**
  - Influence of Submitter
  - Quality of the story
- **Story Quality**
  - Random variable
  - Average out by aggregating fan votes over all stories submitted by the same submitter
  - 289 submitters at least 2 stories
- **Estimate of Influence**
  - Average fan votes
- **Rank Users**

# Statistical Significance of Fan Votes as a Measure of Influence



## URN MODEL



Users in OSN (N)

Fans of submitter in OSN (K)

No. of users who voted (n)

No. of fans who voted (k)

Balls in the urn (N)

White balls in the urn (K)

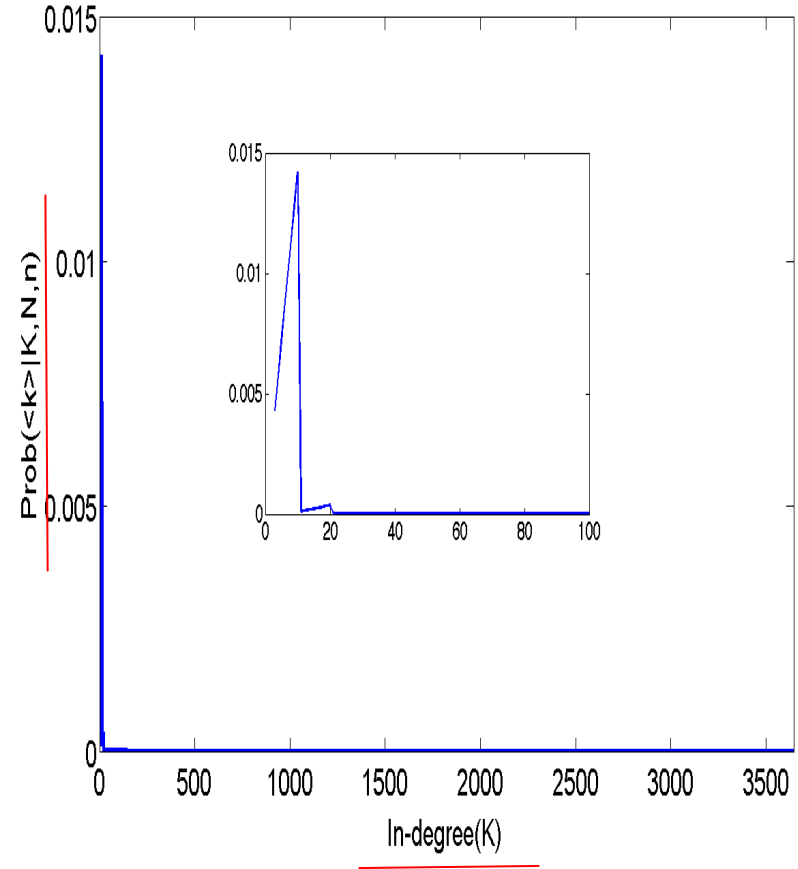
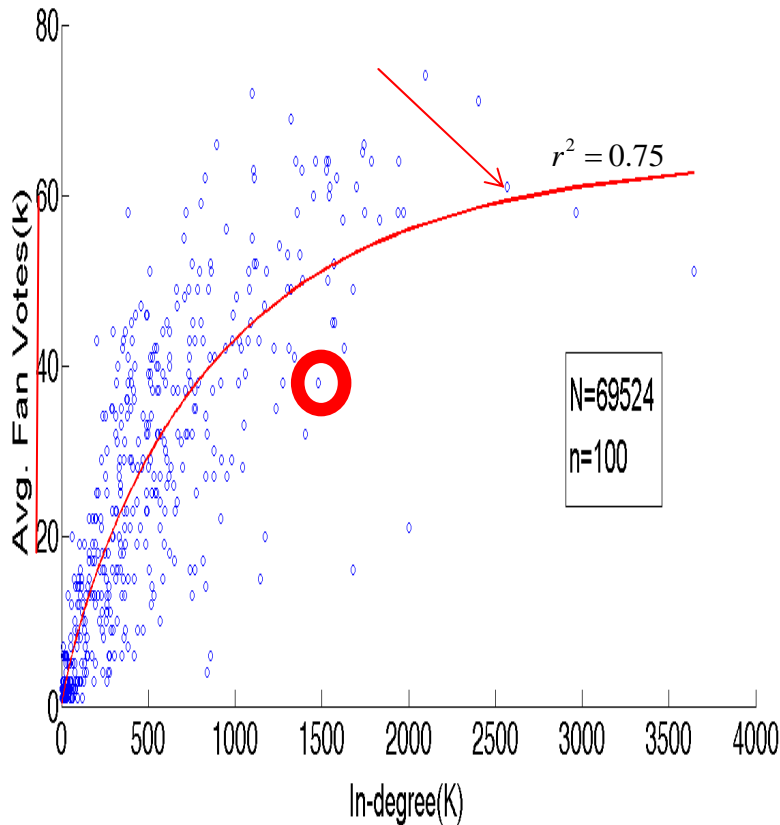
No. of balls picked (n)

No of white balls picked (k)

$$P(X = k | K, N, n) = \frac{\binom{K}{k} \binom{N-k}{n-k}}{\binom{N}{n}}$$

(Hypergeometric Dist.)

# Statistical Significance (Results)



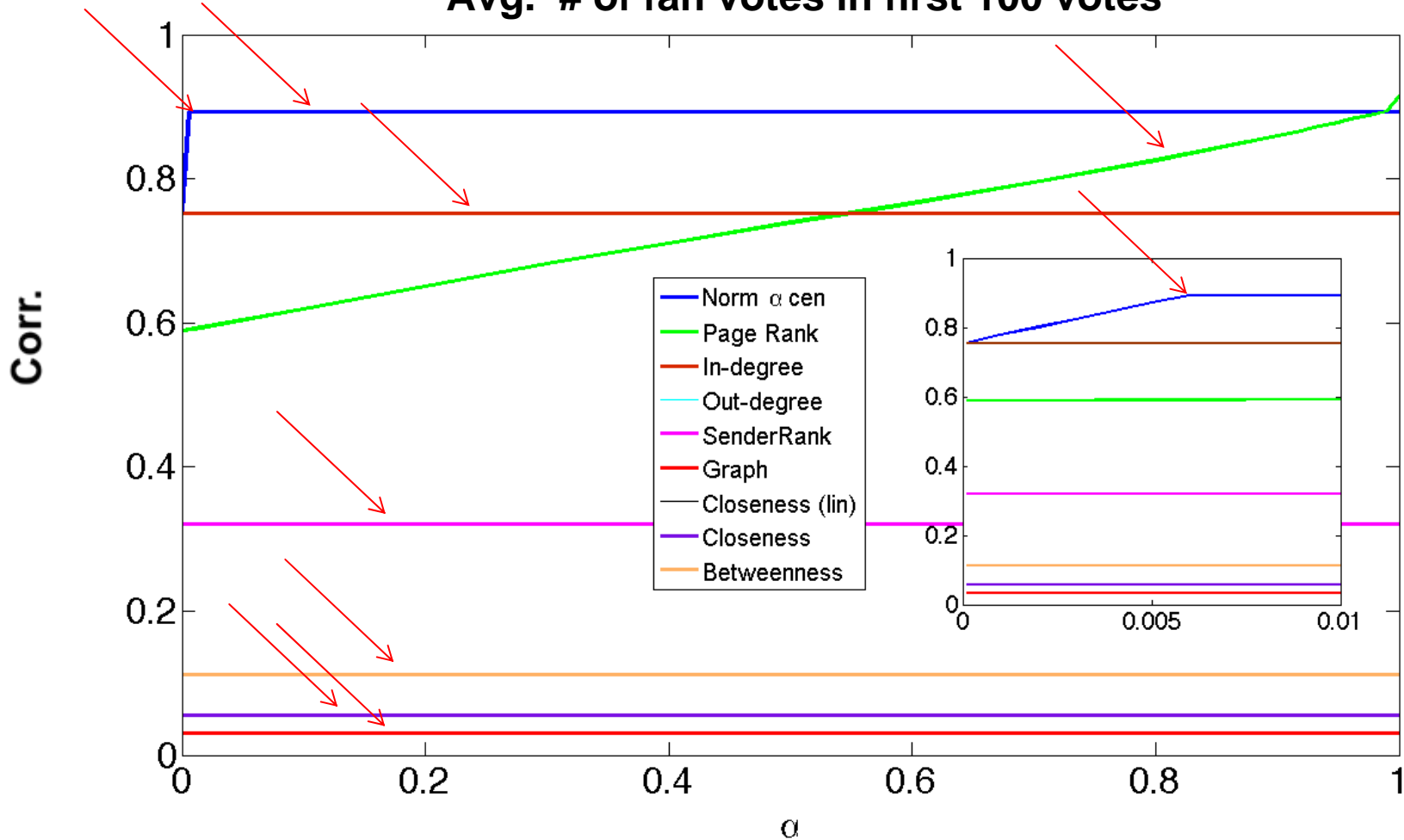
$$\langle k \rangle = 65 \left( 1 - e^{-(0.001 K + 0.0005)} \right)^{0.86}$$

$$P(X = \langle k \rangle | K > 10,69524, 100) < 0.00038$$

**Probability of  $\langle k \rangle$  fan votes in first 100 votes, given the submitter has  $K$  fans, happening purely by chance is negligible**

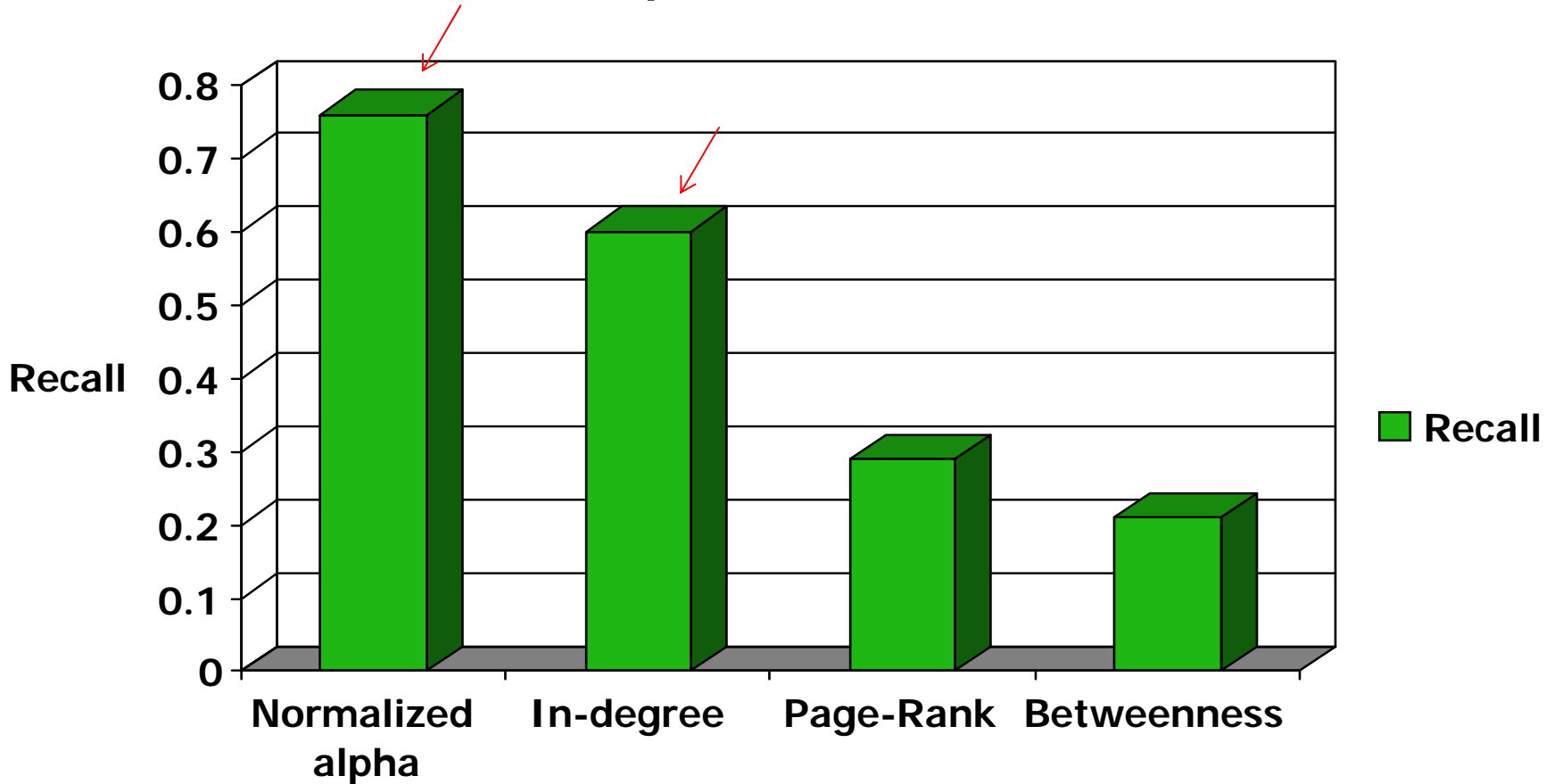
## Correlation with the empirical estimate of influence

Avg. # of fan votes in first 100 votes



$\alpha$ =damping (attenuation factor) in PageRank, (normalized)  $\alpha$ -centrality, SenderRank

## Top 100 users



$$emp(i) \in [1,100]$$

$$pred(i) \in [169,524]$$

$$R = \frac{|emp \cap pred|}{|emp|}$$

- **Results corroborate our hypothesis**

Since underlying **non-conservative** dynamic process

of **(normalized)  $\alpha$ -centrality**

most closely **resembles**

the dynamic process of **information propagation** on Digg

**(normalized)  $\alpha$ -centrality** is a **better predictor** of influential users on Digg than other influence models.

- **How to choose Prediction Models ?**
  - First work classifying influence models into conservative and non-conservative
  - To get the best predictions
    - *choose that influence model whose the implicit dynamic process matches that on the network*
- **How to evaluate Influence Models ?**
  - First work evaluating predictive models of influence using the empirical measurements obtained from the network itself
  - Novel Method of evaluation
    - *Evaluate using influence score estimated empirically from the network*
    - *Social News Aggregator Digg*
    - *Dynamic Process-Information propagation*
    - *Non-conservative influence models best predict influentials on Digg where the underlying dynamic process of information propagation is non-conservative in nature.*
- **Normalized  $\alpha$ -Centrality**
  - Mathematical formulation and analytical proofs