The Evolution of Wikipedia

CS 224W Project by Jiaji Hu, Haozhun Jin, Peng Qi (Group 48)

Introduction

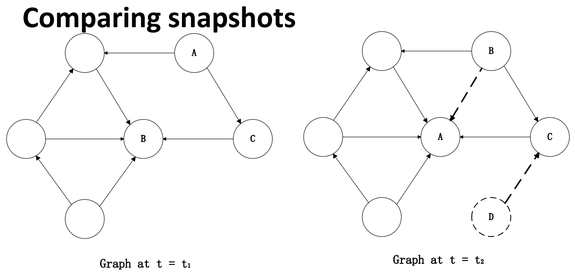
- Social vs Knowledge networks
- Dynamic vs Static viewpoints
- Model proposal: Preferential Attachment by PageRank

Data Collection and Processing

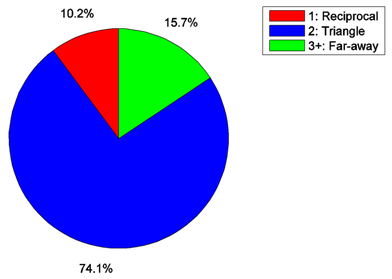
- Dataset
  - Wikipedia edit history before 1/1/2006
- Preprocessing
  - Extract title and link from revision history
  - Sort all revisions by time
  - Take snapshots as necessary
- Special considerations
  - Ignore special pages
  - Remove redirections

Dynamic Analysis

- **Comparing snapshots**

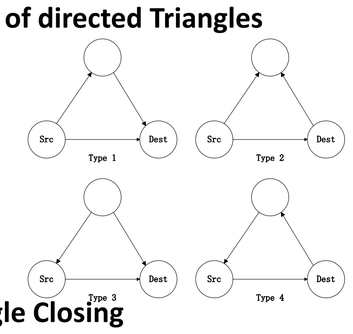


- **Triangle Closing**

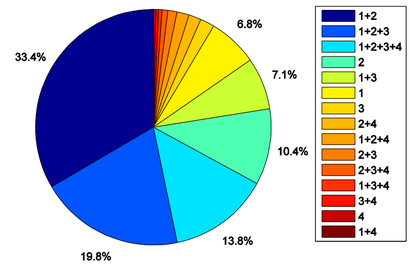


- Majority of new edges close triangles

- **Types of directed Triangles**



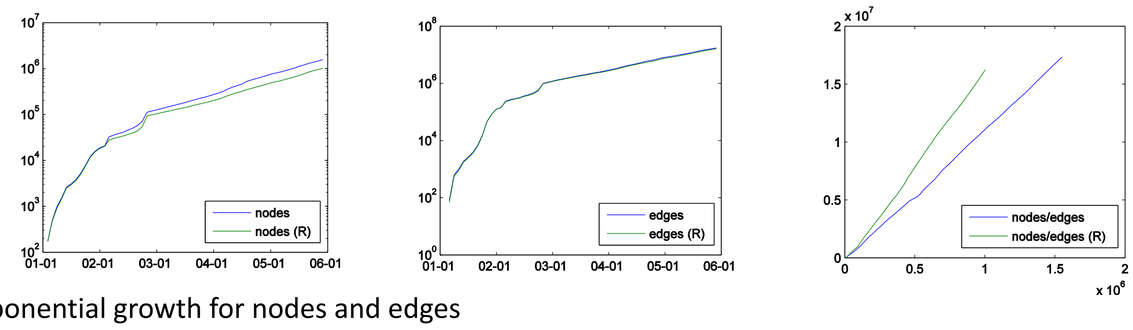
- **Triangle Closing**



- Type 1 and 2 most popular
- Combinations ‘1+2’ and ‘1+3’ most popular

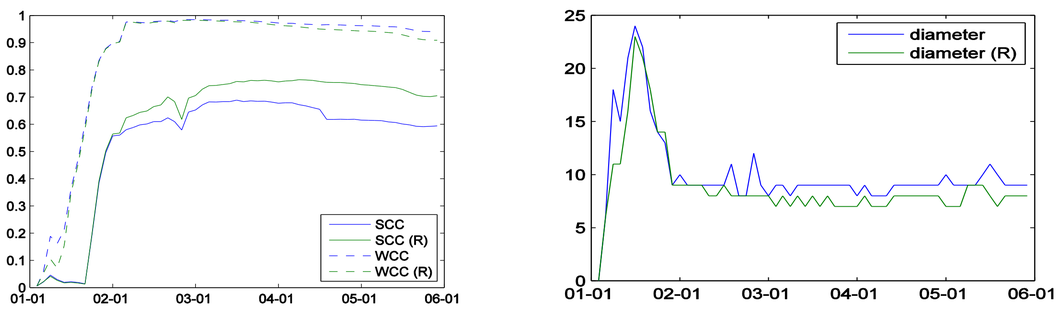
Snapshot Analysis

- **Node & Edge Growth**



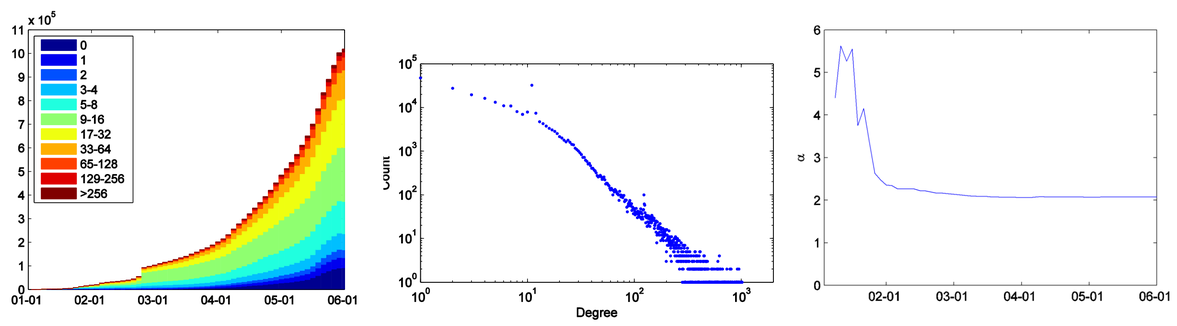
- Exponential growth for nodes and edges
- Roughly linear node/edge ratio

- **Connected Component Sizes & Graph Diameter**



- Removing redirections significantly reduces SCC size
- Removing redirections lower diameter by approx. 1

- **Degree Distribution & Power Law Exponent**

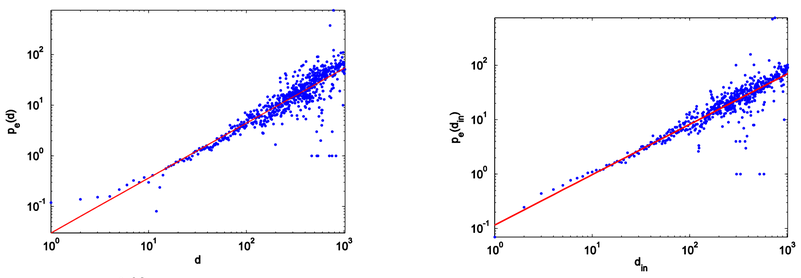


- Power law degree distribution
- Exponent α converges to approx. 2

Model Proposal

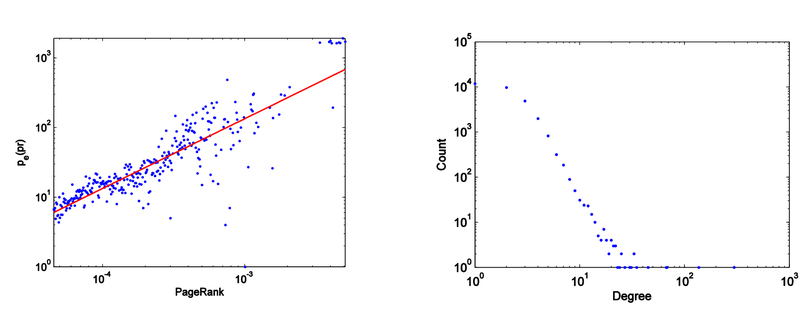
Edge Destination Prediction

- **Preferential Attachment**



- $p_e(d) = d^{1.08}$
- $p_e(d_{\text{in}}) = d_{\text{in}}^{0.92}$
- Roughly linear attachment by degree & in-degree

- **Preferential Attachment by PageRank**



- $p_e(pr) = pr^{1.00}$
- PA-PageRank generation model creates scale-free network

Conclusion

- Static and dynamic analysis of Wikipedia network.
- Snapshot analysis indicates Wikipedia is a scale-free network similar to social networks.
- Dynamic analysis confirms previous triangle closing observations. Types of closed directed triangles studied with intuitive findings.
- Preferential attachment by degree confirmed on Wikipedia. Preferential attachment by PageRank proposed and confirmed.
- PA-PageRank generation model proposed and implemented.