# TweetRank

Shayon Gupta & Gerard Casas Saez
Design and Analysis of Algorithms class
University of Colorado Boulder

# PageRank

# PageRank

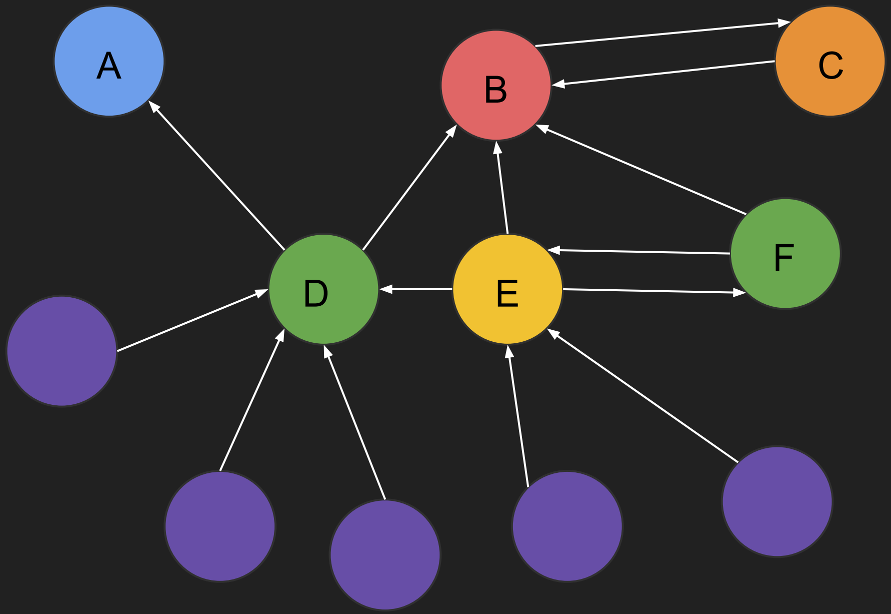
PageRank is a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.

Sergey Brin and Larry Page, 1998

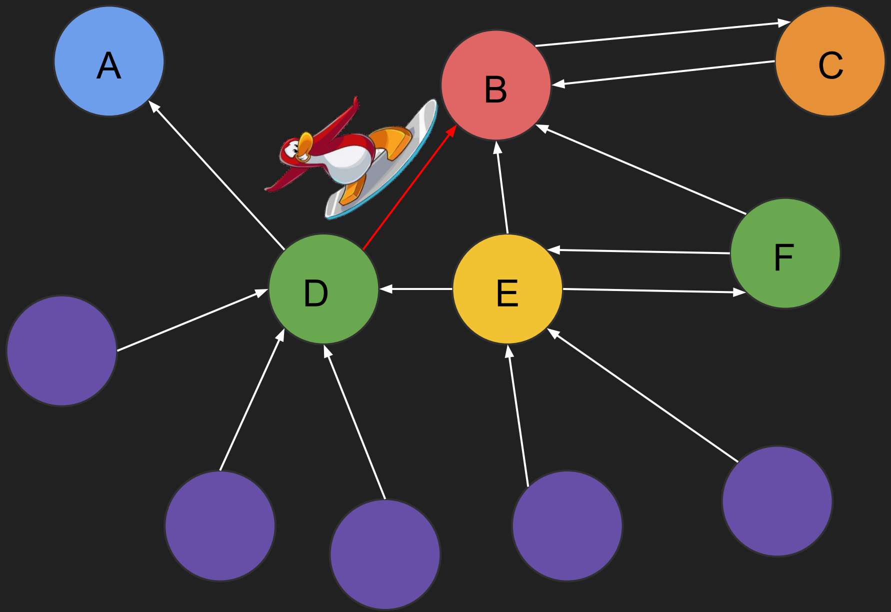
# **PageRank**: The Origin

- **1976**: Eigenvalue problem proposed by Gabriel Pinski and Francis Narin
- Sergey Brin has the idea to order web data in a hierarchy by "**link popularity**"
- **1998**: PageRank Citation Ranking: Bringing order to the internet
- **1998**: Google paper
- **1998**: Google is founded by Sergey Brin and Larry Page
- Named after Larry Page and web page
- Patented assigned to **Stanford** University, not Google

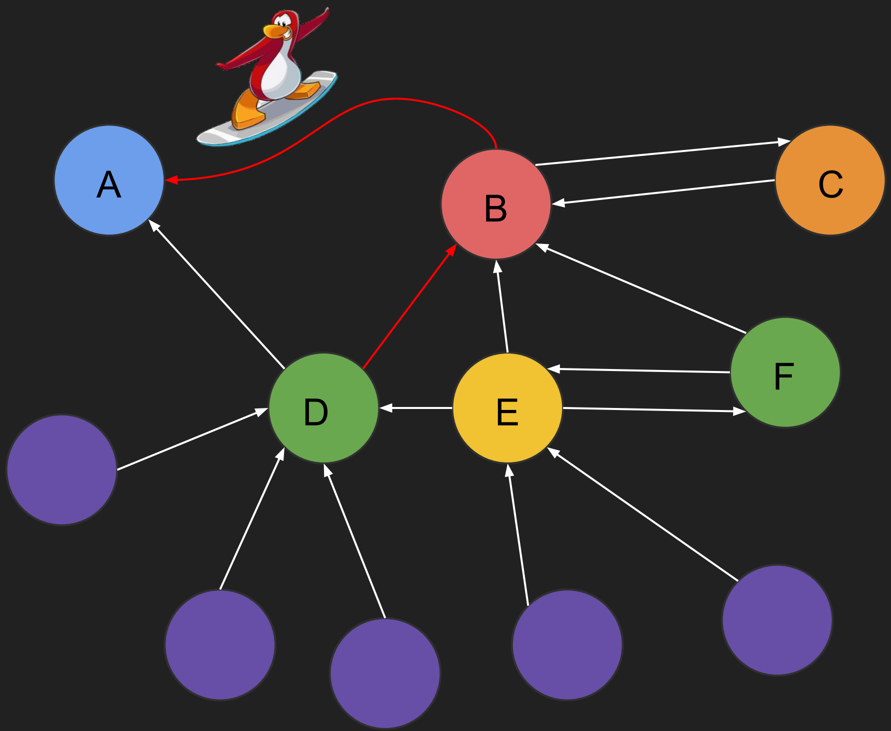
# **PageRank**: How?



# **PageRank**: Random surfer model



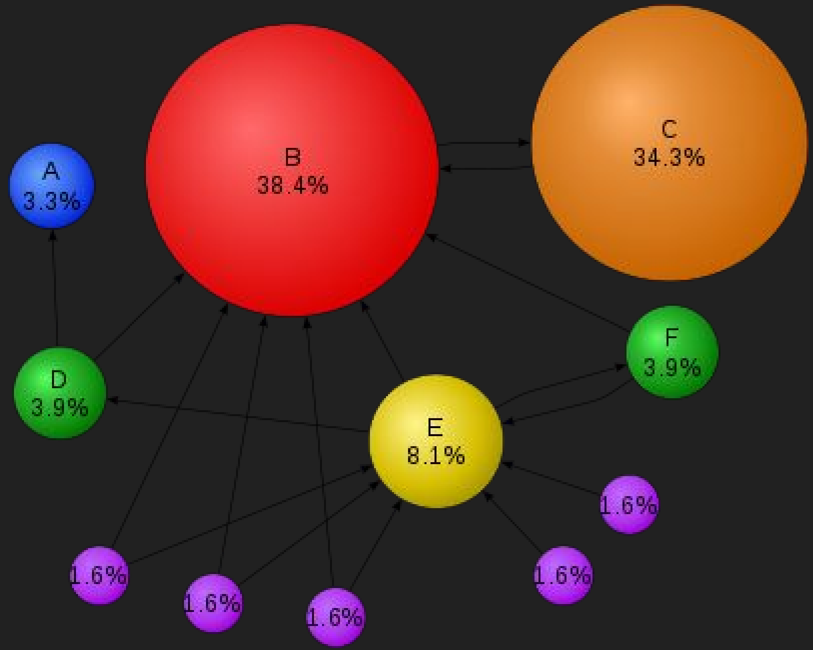
# **PageRank**: Damping factor



# **PageRank**: Real formula

$$PR(B) = \frac{1-d}{N} + d\left(\frac{PR(D)}{L(D)} + \frac{PR(E)}{L(E)} + \frac{PR(F)}{L(F)} + \frac{PR(C)}{L(C)}\right)$$

# **PageRank**: Result




Credits to Wikipedia for this graph

# Dataset

# Twitter Data

- Twitter data **openly** available for research
- Public REST API
- Returns: list of tweets and retweets
- Relational data:
  - **Retweets**: Publish other users tweets in your own timeline
  - **Likes**: User likes content
  - **Follow/Followed**: Aggregate users data to see in home timeline

# Dataset

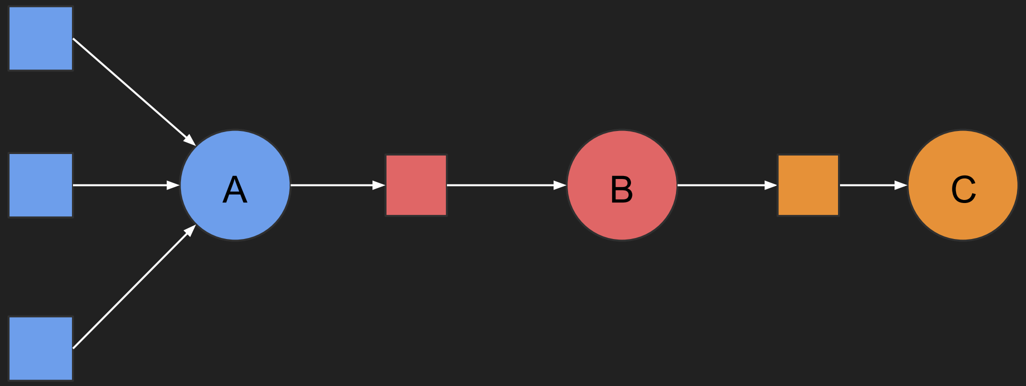
- **18215** tweets and retweets.
- **4197** tweets
- Collected at Project EPIC
    - Analysis of tweets generated during crisis to improve public response
- Tweets about hurricanes **Harvey and Irma**

# TweetRank

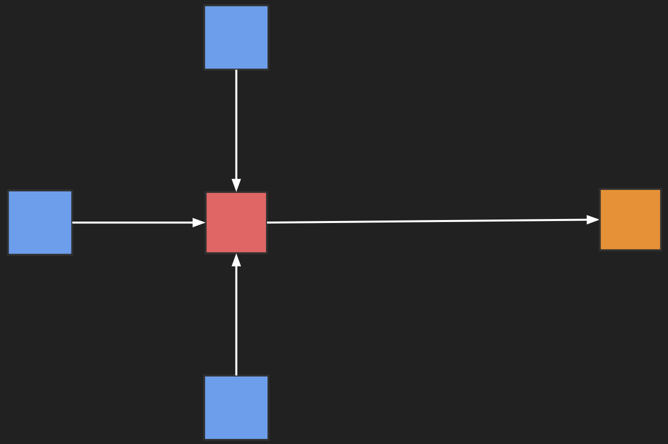
# **TweetRank**: Motivation

- Users in disasters usually retweet tweets in the **same disaster area** [Marina Kogan et al. 2015]
- Tweets locally retweeted usually have **local utility** [Marina Kogan et al. 2015]
- **Faster** than news outlets (or similar)
- Interesting for **disaster emergency response on site**

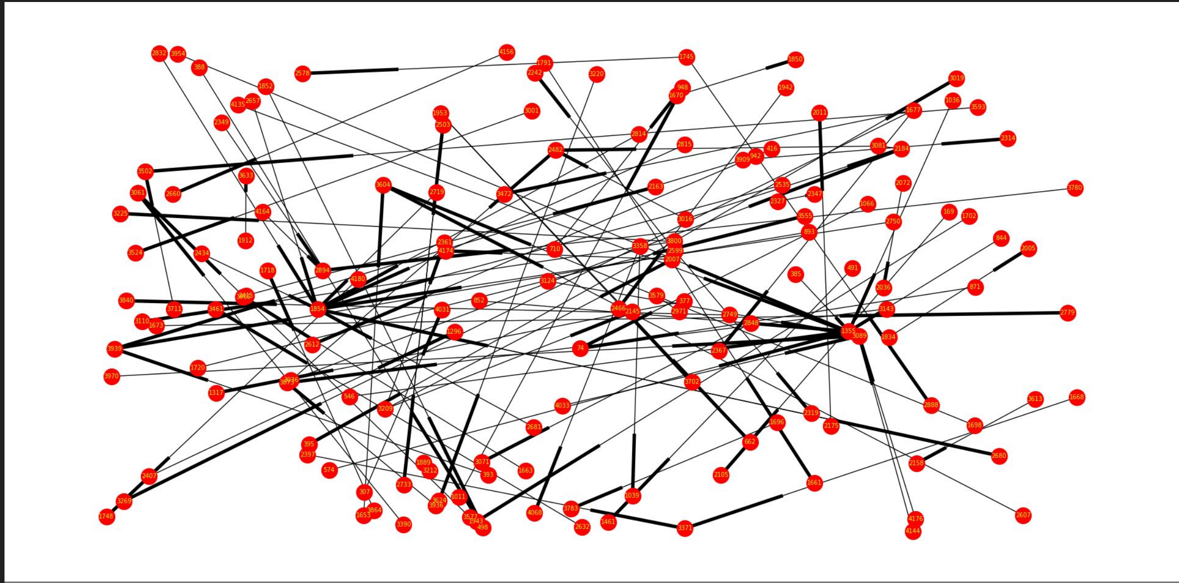
# **TweetRank**: Random twitter user model



# **TweetRank**: Random twitter user model



# **TweetRank**: Graph Visualization



This Di-Graph shows the relationships between the tweets and their retweets.

# **TweetRank**: Implementation

- Coded in Python
- 2 main parts
  - Graph generation (json, dictionaries)
  - Algorithm execution (Panda dataframes)
- Not really scalable, but can be easily updated to Spark Dataframes

# **TweetRank**: Results

- Damping factor: **0.85**
- Error below **0.0001**
- Convergence in **3 iterations** for dataset

```
Running TweetRank...done
904814761141030912    0.000238
904813816965459969    0.000238
904815764733120512    0.000157
903201962744836096    0.000153
904813409136463879    0.000127
904811943755644928    0.000096
904815149617426433    0.000096
904814132888834048    0.000096
904557396340637696    0.000096
904704308838625282    0.000072
```

# **TweetRank**: Top10 sample

# **TweetRank**: "Interpretation"

- Tweets ranked by importance **inside** dataset
- If user has more tweets in dataset, user's retweets more "**meaningful**"
- More to come?

# **TweetRank**: Future

- More research
    - Check results **interpretation**
    - Study **scalability**
    - Compare with other methods
- Run algorithm in **bigger** datasets
- **Locally geo bounded** datasets [Marina Kogan et al. 2015]

# References


- PageRank Citation Ranking: Bringing order to the internet [Larry Page et al. 1998]
- The Anatomy of a Large-Scale Hypertextual Web Search Engine [Sergey Brin and Larry Page. 1998]
- Think Local, Retweet Global: Retweeting by the Geographically-Vulnerable during Hurricane Sandy [Marina Kogan et al. 2015]

Questions?

Thanks! 👍

# **TweetRank**: Graph

106 edges

N tweets with an edge

# **PageRank**: Time analysis

- 322 million links -> 52 iterations
- 161 million links -> 45 iterations until similar convergence
- Logarithmic time on the size of the graph if the graph is expander

# **PageRank**: Formula in the original paper

$$PR(B) = {}_{1-d} + d\left(\frac{PR(D)}{L(D)} + \frac{PR(E)}{L(E)} + \frac{PR(F)}{L(F)} + \frac{PR(C)}{L(C)}\right)$$