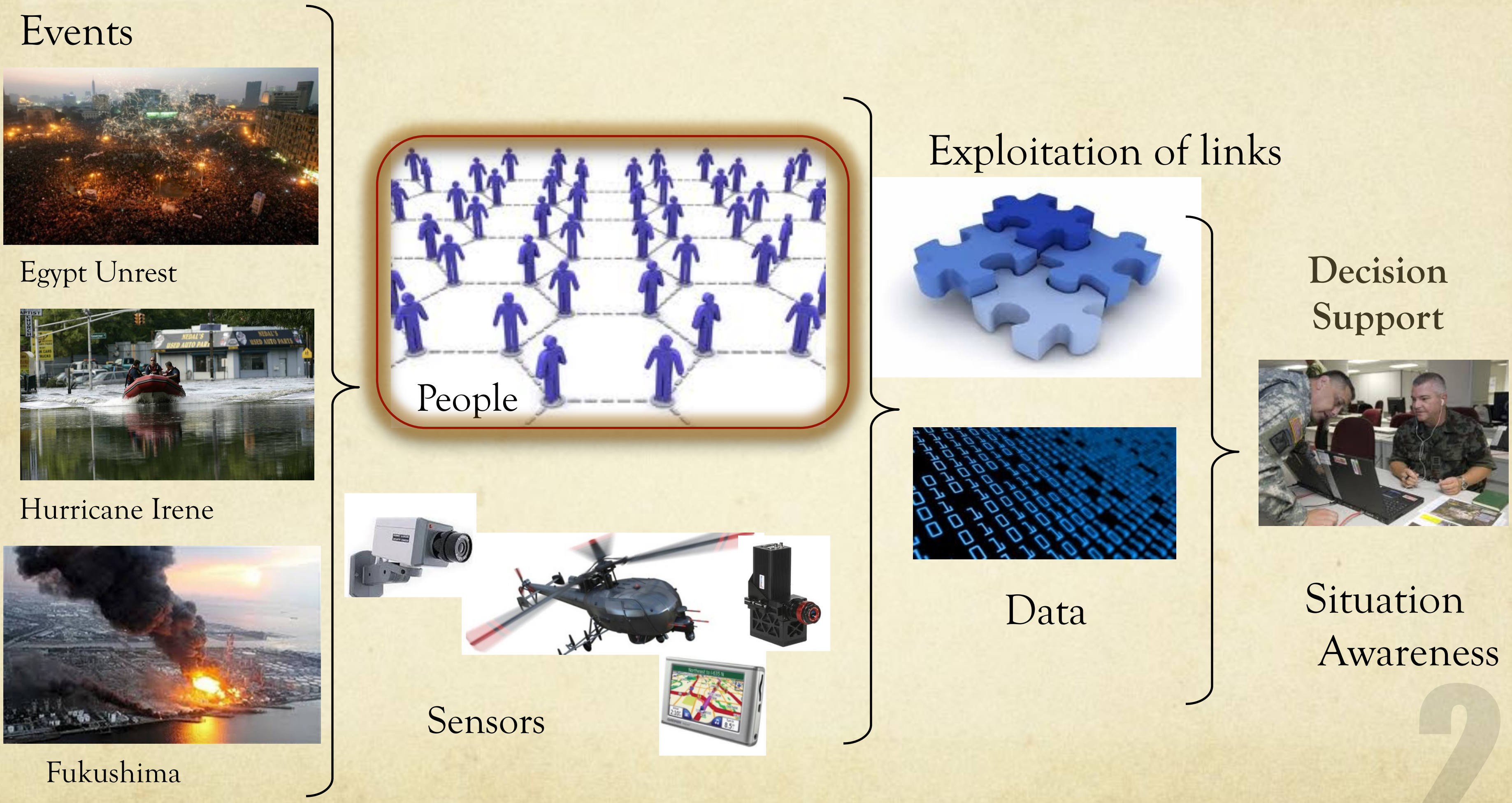
# On Diversifying Source Selection in Social Sensing

Md Yusuf Sarwar Uddin, Md Tanvir Al Amin, Hieu Le, Tarek Abdelzaher, Boleslaw Szymanski, Tommy Nguyen

Presented By: Md Tanvir Al Amin

# An Information Pipeline



Events

Egypt Unrest

Hurricane Irene

Fukushima

People

Sensors

Exploitation of links

Data

Decision Support

Situation Awareness

Portions of slide taken from “Information Analysis of Social and Information Links”

# People as Sensors

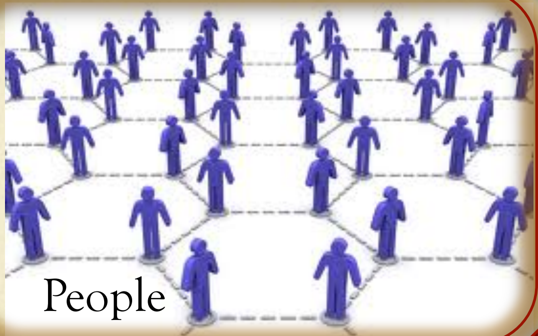
## A New Kind of Social Sensing

Events

Egypt Unrest

Hurricane Irene

Fukushima



People

Twitter

(140 Character messages)

Ahmed: I saw a lot of people gather in Tahrir Square

Rahman: Gunfire sounds in Tahrir Square

Flickr

(Photo Sharing)

Facebook

(Social Networking)

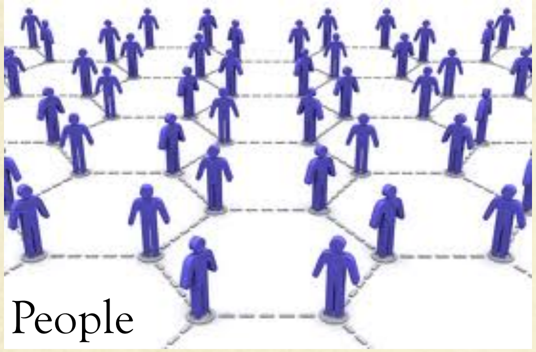
# Fact Finders

Events

Egypt Unrest

Hurricane Irene

Fukushima



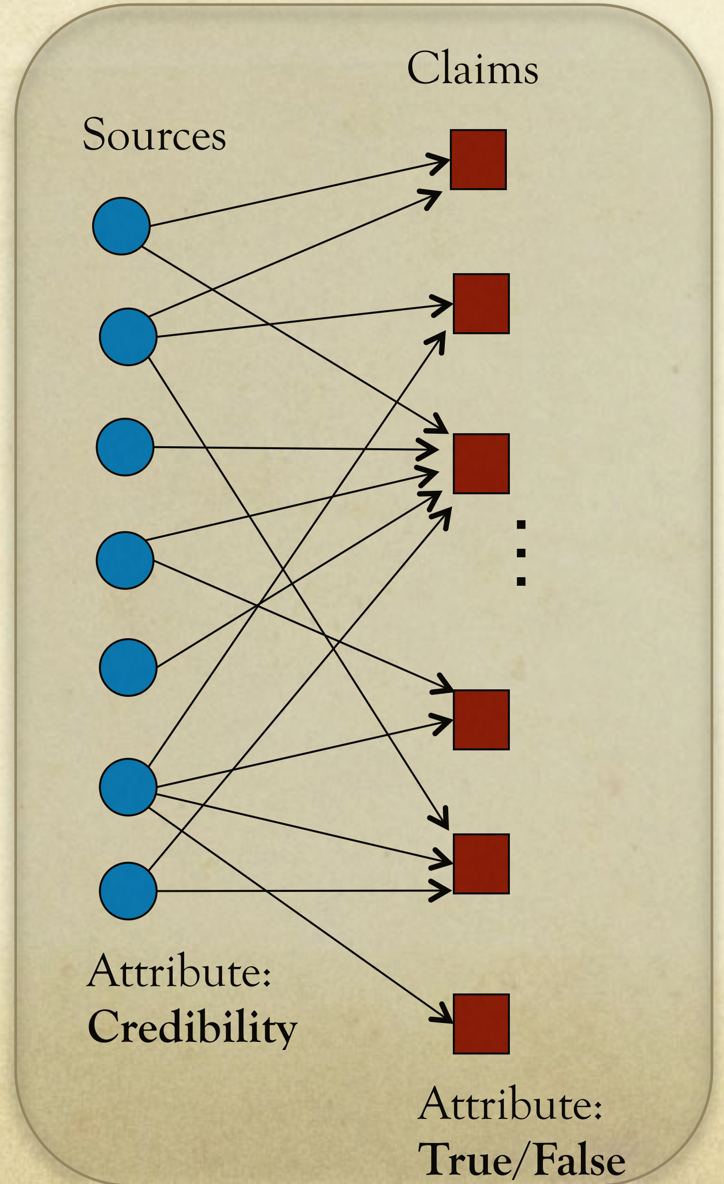
Real-time Reports (Example: Tweets)

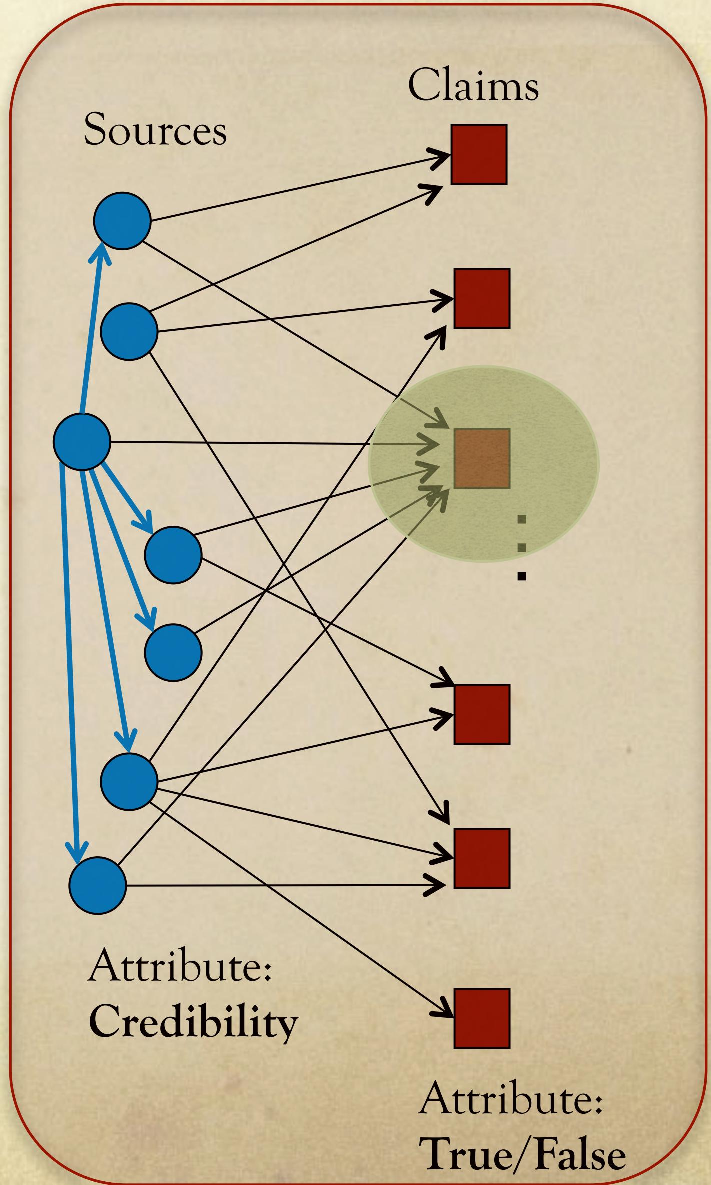
Reconstruct event timeline (what really happened?)

# Fact Finders State of the Art

- Hubs and Authorities - 1999
- TruthFinder - 2008
- 3-Estimates - 2010
- AccuVote – 2010
- Pasternack et. al – 2010
- Gupta et. al. – 2011
- Apollo - 2011

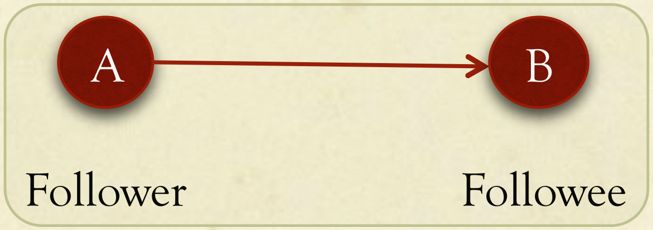
# Source Dependency



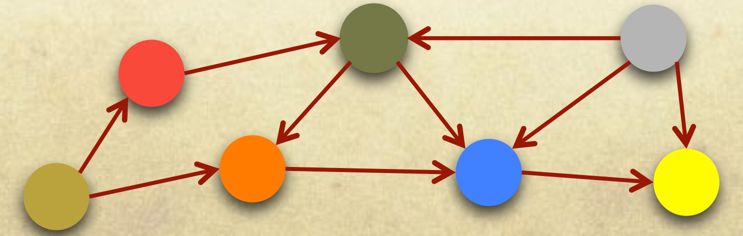


# Assessing Source Dependency

- Source dependency can be modeled with the aid of the social network among the users.
- For example, in Twitter, user A can *follow* another user B, which means A has subscribed to receive the updates of B.

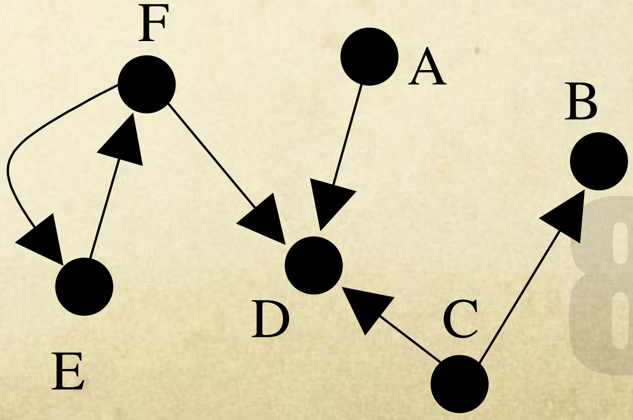


- Set of this follower-followee relationships create a network which forms a *Social Graph*



# Source Selection

- Have a *distance metric* between source pairs, that can be
  - Function of their shortest path length in the *social graph*
  - Function of their geographic distance
  - Function of number of common followers or followees
  - May be something else …
- Formally distance is $1 - f_{ij}$ where $f_{ij}$ is a *dependency function* between $i$ and $j$
  - With probability $f_{ij}$, source $i$ could make the same or similar claims as source $j$



# Formal Statement

- $V$ is the set of all sources, $S$ is the set of selected sources

- *Independence Score* $\beta(i,S)$ for each of the sources $i$ in $S$ is a *measure of its independence* in making claims, with respect to the other selected sources

$$\beta(i,S) = \prod_{j \in S} (1 - f_{ij})$$

- Find $S$ so that the *Sum of Independence Scores* over $S$ is maximized

$$\boxed{\begin{array}{l} \max \sum_{i \in S} \beta(i,S) \\ \text{subject to } \beta(i,S) \geq \tau, \forall i \in S \end{array}}$$

$$= \max \sum_{i \in S} \prod_{j \in S} (1 - f_{ij})$$

$$\text{subject to } \prod_{j \in S} (1 - f_{ij}) \geq \tau, \forall i \in S$$

# Does it Scale?

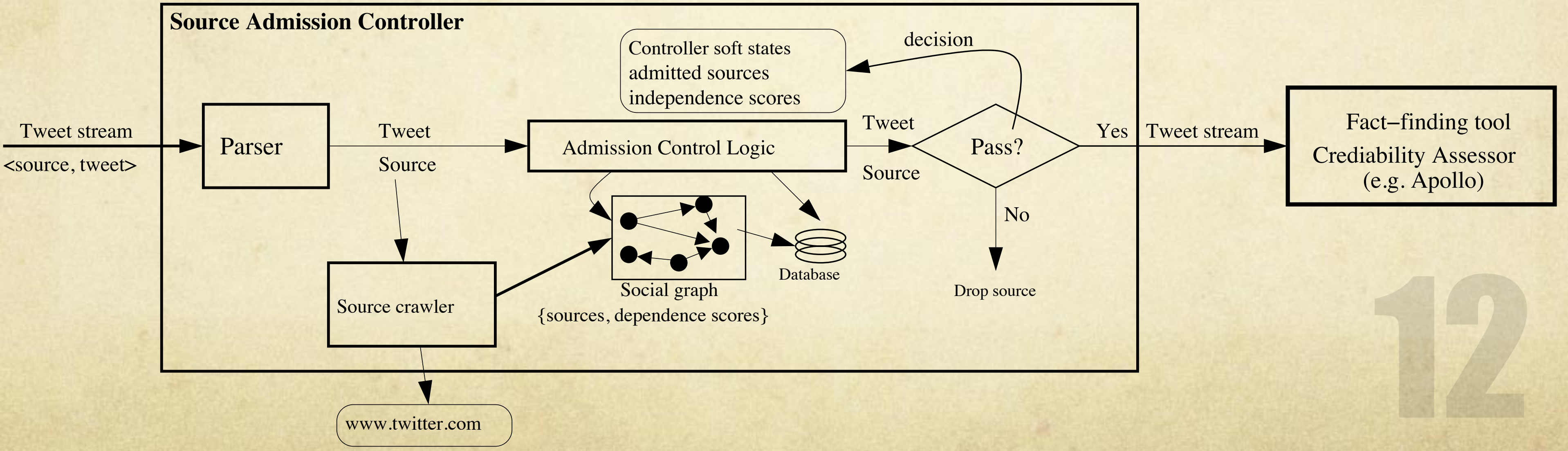
Photo courtesy: http://www.midatlanticoceanresearchplan.org/offshore-energy-development

# Computing a Solution

- Tweets arrive in real-time, like the streams.
  - Never know who is going to tweet next !
- Its not practical to crawl the whole of social network among all the users beforehand
  - Too large number of sources!
  - Problem itself NP-Hard

# Schematic Model

- Go for an online solution – Greedy Algorithm
  - Prefix an *Admission Controller* to *Apollo* pipeline
- *Admission Controller* passes or rejects tweets according to available information
  - Set S is computed incrementally

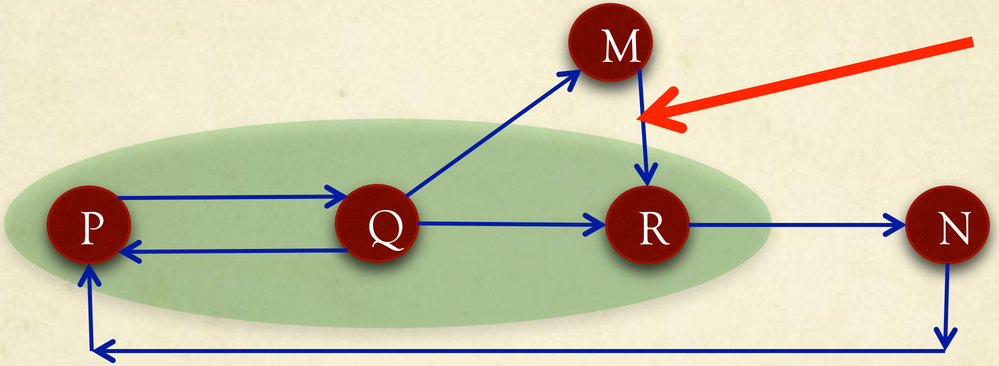


# Admission Control Schemes

- By defining the *dependency function* and the *threshold* appropriately, different admission controllers can be achieved.
- For example:
  - *No Direct Follower*
  - *No Direct Follower + No Common Followee.*
  - *No Descendants*
  - *$\beta$ - Controller*

# No Direct Follower

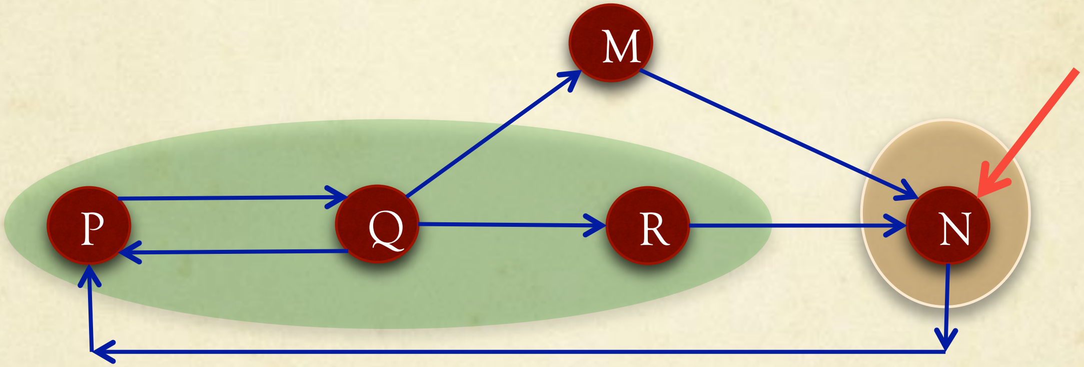
- Deny if the source is a direct follower of another already admitted source.



- S = {P, Q, R} and M is a new source
- M rejected because it follows R already in S

# No Direct Follower + No Common Followee

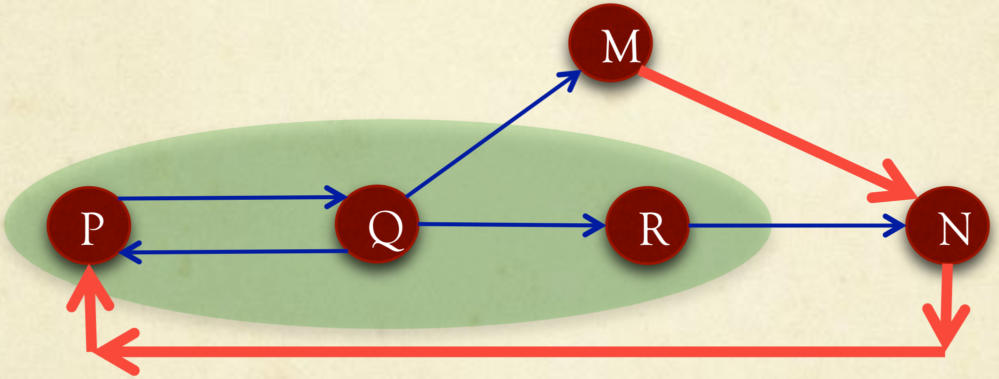
- Deny if the previous condition holds or the source has at least one common followee with another admitted source



- *S* = {*P*, *Q*, *R*} and M is a new source
- M rejected because it follows N and R also follows N

# No Descendents

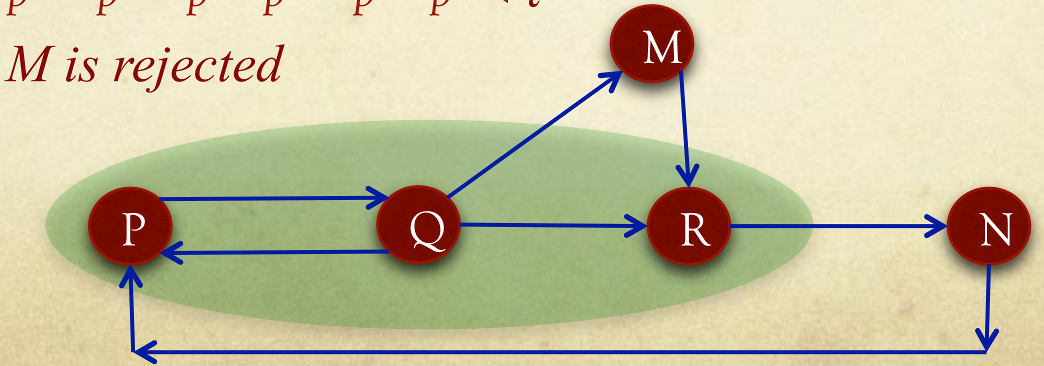
- Deny if the source is a follower of another admitted source possibly via a set of intermediate followees.



- *S* = {*P*, *Q*, *R*} and M is a new source
- M rejected because it follows P through a chain

# $\beta$ - Controller

- At each step, select the source if it improves *Independence Score* of the set S by an amount of at least $\tau$
  - Dependency function $f_{ij}$ taken to be $p^k$, where $k$ is the length of path from $i$ to $j$. $p$ is a "information flow" probability from 0 to 1.
  - Therefore, if S = {$P$, $Q$, $R$} and M is a new source
    - $f_{MP} = p^3$, $f_{MQ} = p^4$, $f_{MR} = p$,
    - $f_{PM} = p^2$, $f_{QM} = p$, $f_{RM} = p^4$,
    - $p^3 + p^4 + p - p^2 - p - p^4 < \tau$
    - *M is rejected*

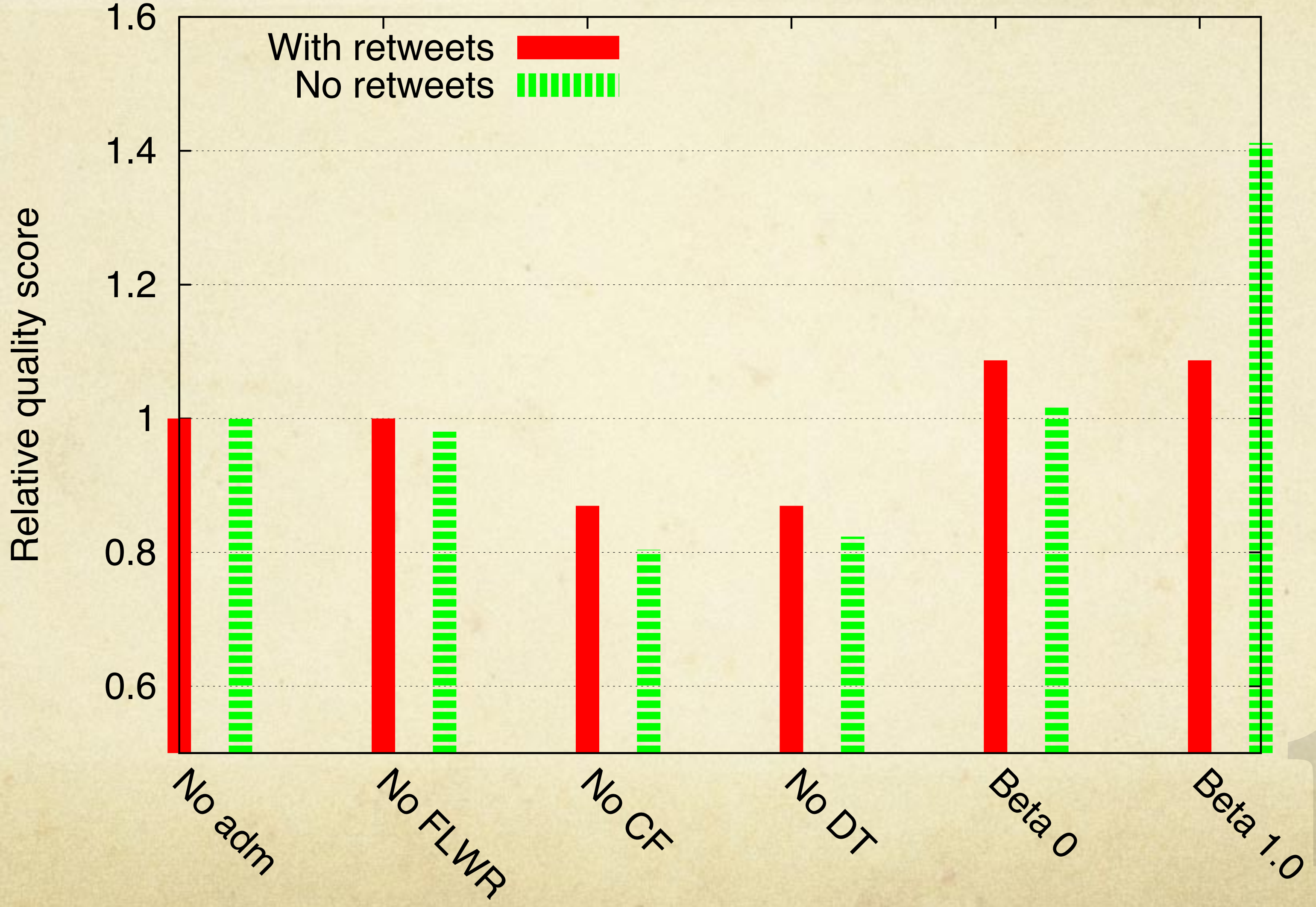


# Evaluation

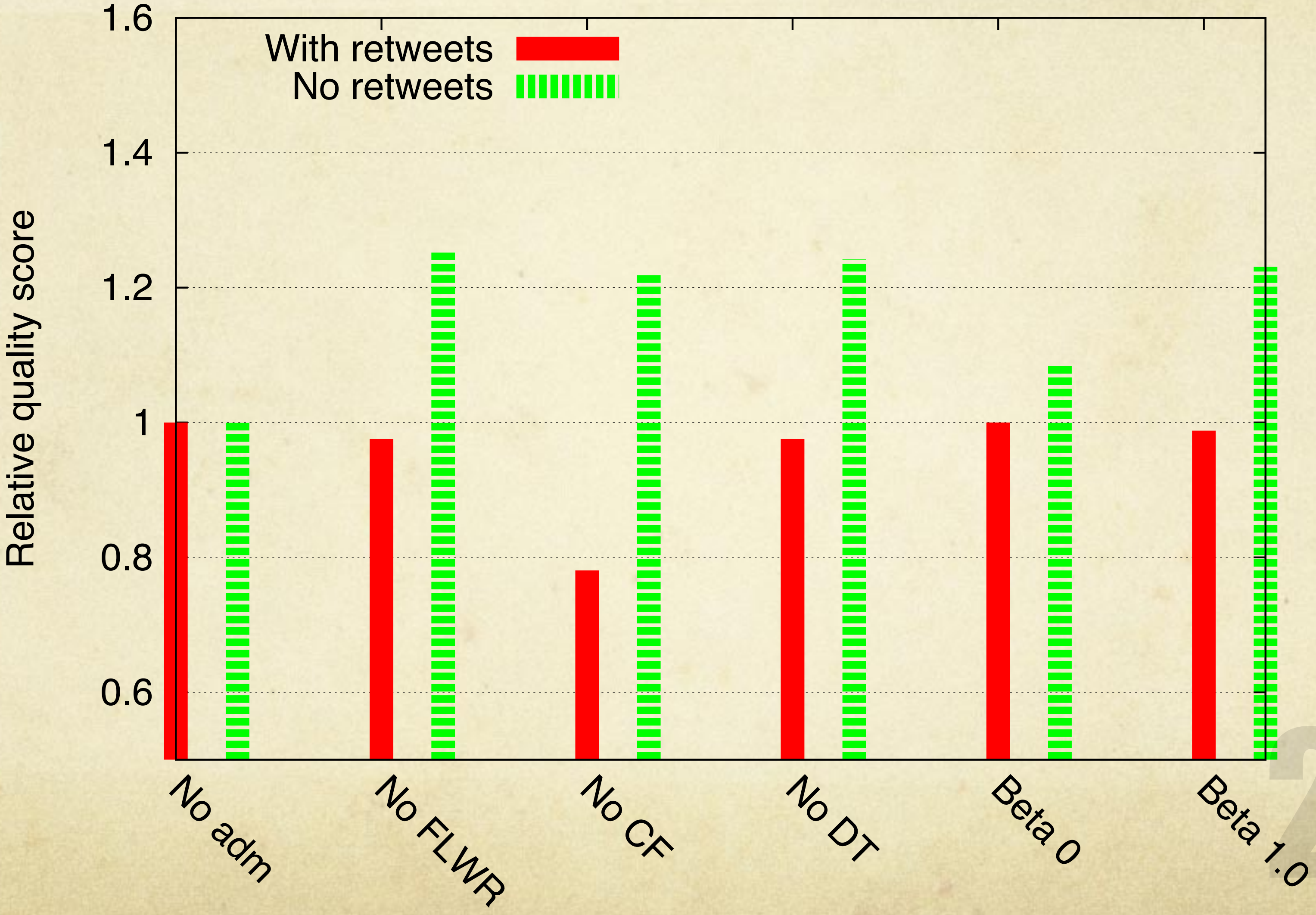
- Evaluations done on two datasets
  - Egypt Unrest (dense dataset)
  - Hurricane Irene (sparse dataset)

| Dataset | Egypt Unrest | Hurricane Irene |
|---|---|---|
| Time Duration | 18 days | 7 days |
| # of tweets | 1,873,613 | 387,827 |
| # of users crawled | 5,285,160 | 2,510,316 |
| # of users actually tweeted | 305,240 | 261,482 |
| # of follower-followee links | 10,490,098 | 3,902,713 |

# Comparative Scores - Egypt



# Comparative Scores – Irene



# Lessons Learned

- Human generated un-vetted data can be noisy, incomplete and misleading. *Dependency and Social Connection* between sources play an important role in the quality of data fusion.
- *Diversifying Source Selection* can improve the quality of fact finders.
- Experiment shows that *beta admission control p*erforms best in general.

# Conclusions

- Socially obtained data is not independent in general, we have suggested to consider the social network to select only a subset of the sources.
- We have mathematically formulated the above problem for optimality.
- We have provided a customizable online algorithm to perform the source selection in real-time, in amortized O(1) time. We have generated four heuristics from that algorithm.
- Experimental results say that source selection is necessary to improve the quality of data fusion.

# Discussion

# Discussion Questions

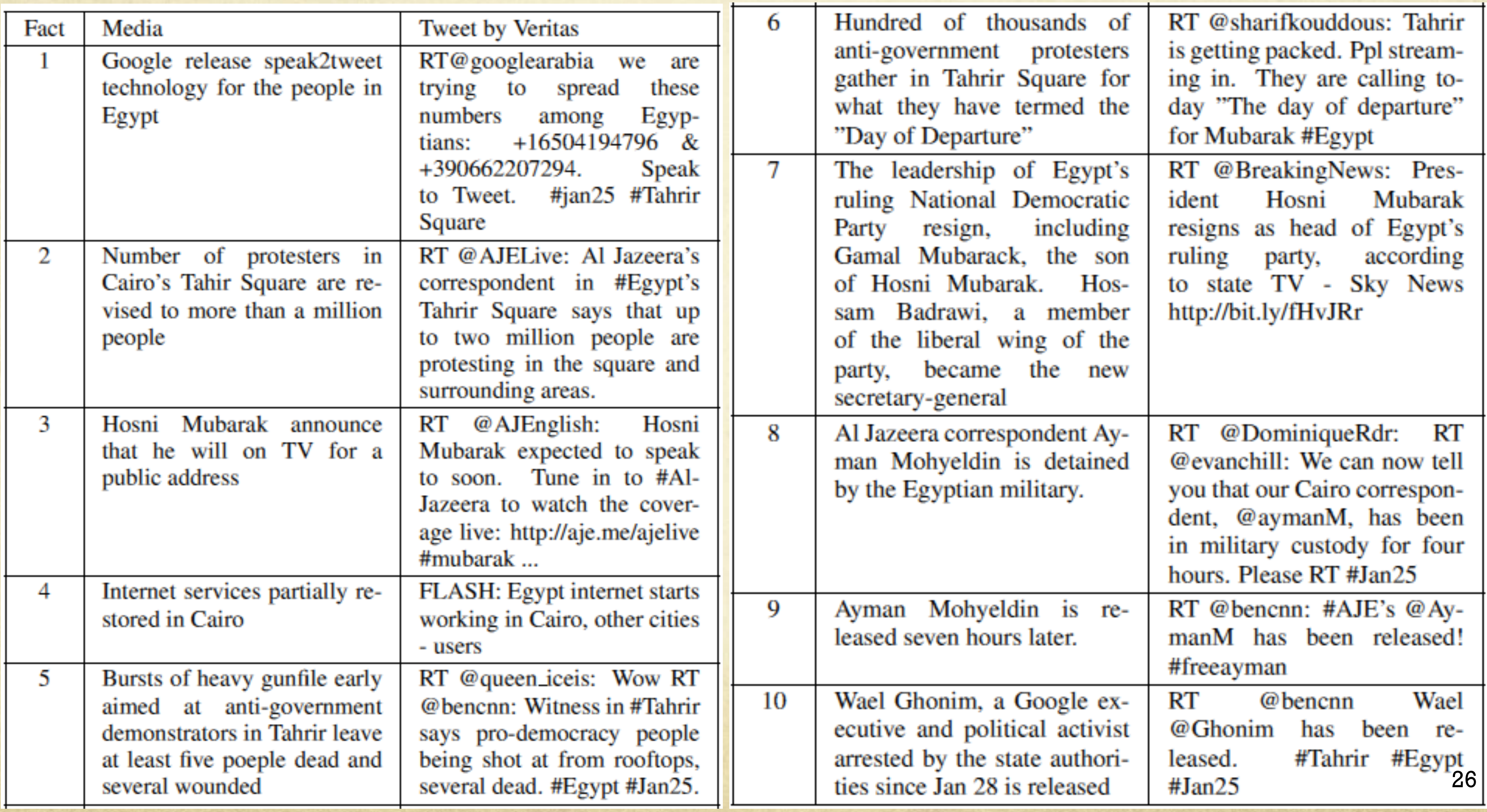
- Why the admission control verdicts on *sources*? Isn't it be more logical to decide on *tweets* instead?
- Why the admission controller is remembering its decisions? Shouldn't it periodically re-asses the admissibility of the sources?

# Backup Slides

# “Most Credible” Tweets Egypt Uprising

- **Example: Summarizing Twitter Feeds**

  1.5 Million tweets collected during Egypt Uprising (Feb/March 2011). Examples of “top tweets” from produced event summary and corresponding media reports



| Fact | Media | Tweet by Veritas |
|---|---|---|
| 1 | Google release speak2tweet technology for the people in Egypt | RT@googlearabia we are trying to spread these numbers among Egyptians: +16504194796 & +390662207294. Speak to Tweet. #jan25 #Tahrir Square |
| 2 | Number of protesters in Cairo’s Tahir Square are revised to more than a million people | RT @AJELive: Al Jazeera’s correspondent in #Egypt’s Tahrir Square says that up to two million people are protesting in the square and surrounding areas. |
| 3 | Hosni Mubarak announce that he will on TV for a public address | RT @AJEnglish: Hosni Mubarak expected to speak to soon. Tune in to #AlJazeera to watch the coverage live: http://aje.me/ajelive #mubarak ... |
| 4 | Internet services partially restored in Cairo | FLASH: Egypt internet starts working in Cairo, other cities - users |
| 5 | Bursts of heavy gunfile early aimed at anti-government demonstrators in Tahrir leave at least five poeple dead and several wounded | RT @queen_iceis: Wow RT @bencnn: Witness in #Tahrir says pro-democracy people being shot at from rooftops, several dead. #Egypt #Jan25. |
| 6 | Hundred of thousands of anti-government protesters gather in Tahrir Square for what they have termed the ”Day of Departure” | RT @sharifkouddous: Tahrir is getting packed. Ppl streaming in. They are calling today ”The day of departure” for Mubarak #Egypt |
| 7 | The leadership of Egypt’s ruling National Democratic Party resign, including Gamal Mubarack, the son of Hosni Mubarak. Hossam Badrawi, a member of the liberal wing of the party, became the new secretary-general | RT @BreakingNews: President Hosni Mubarak resigns as head of Egypt’s ruling party, according to state TV - Sky News http://bit.ly/fHvJRr |
| 8 | Al Jazeera correspondent Ayman Mohyeldin is detained by the Egyptian military. | RT @DominiqueRdr: RT @evanchill: We can now tell you that our Cairo correspondent, @aymanM, has been in military custody for four hours. Please RT #Jan25 |
| 9 | Ayman Mohyeldin is released seven hours later. | RT @bencnn: #AJE’s @AymanM has been released! #freeayman |
| 10 | Wael Ghonim, a Google executive and political activist arrested by the state authorities since Jan 28 is released | RT @bencnn Wael @Ghonim has been released. #Tahrir #Egypt #Jan25 |

# Independence Score

- *Independence Score* $\beta(i,S)$ for each of the sources $i$ in $S$ is a *measure of its independence* in making claims, with respect to the other selected sources

$$\begin{aligned}&\beta(i,S)\\&= P[i \text{ is independent in making claims}]\\&= \prod_{j\in S} P[i \text{ is not dependent on } j]\\&= \prod_{j\in S}(1-f_{ij})\end{aligned}$$

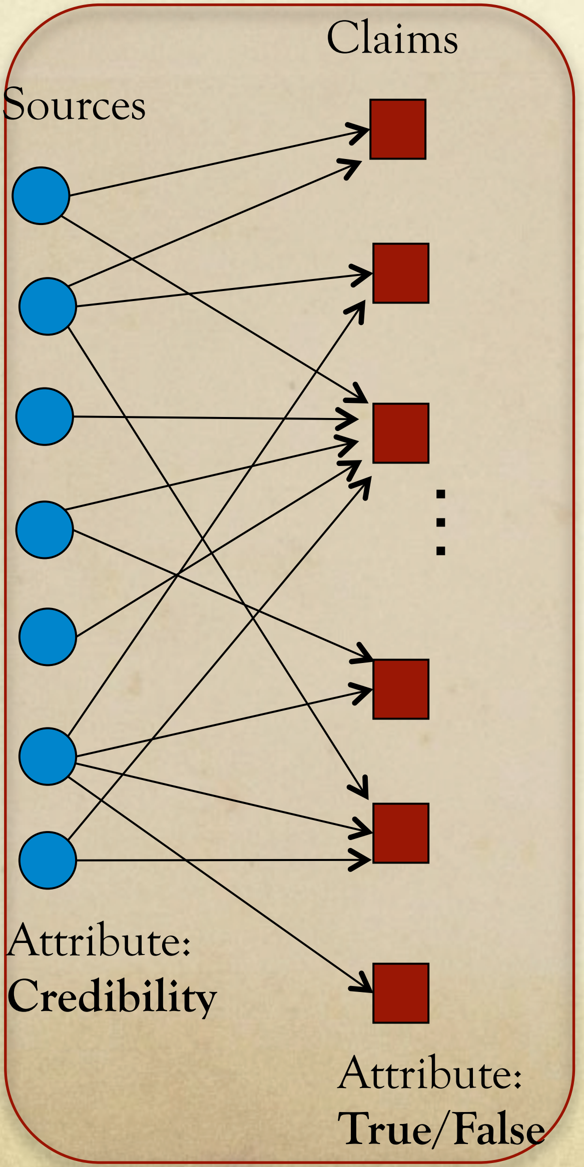
# Apollo Fact Finder

Events

Egypt Unrest

Hurricane Irene

Fukushima



Claims

Sources

Attribute:
**Credibility**

Attribute:
**True/False**

**Maximum Likelihood Estimation**

**Event Summary**

- *Credibility of sources*
- *Correctness of claims*
- *Confidence intervals*

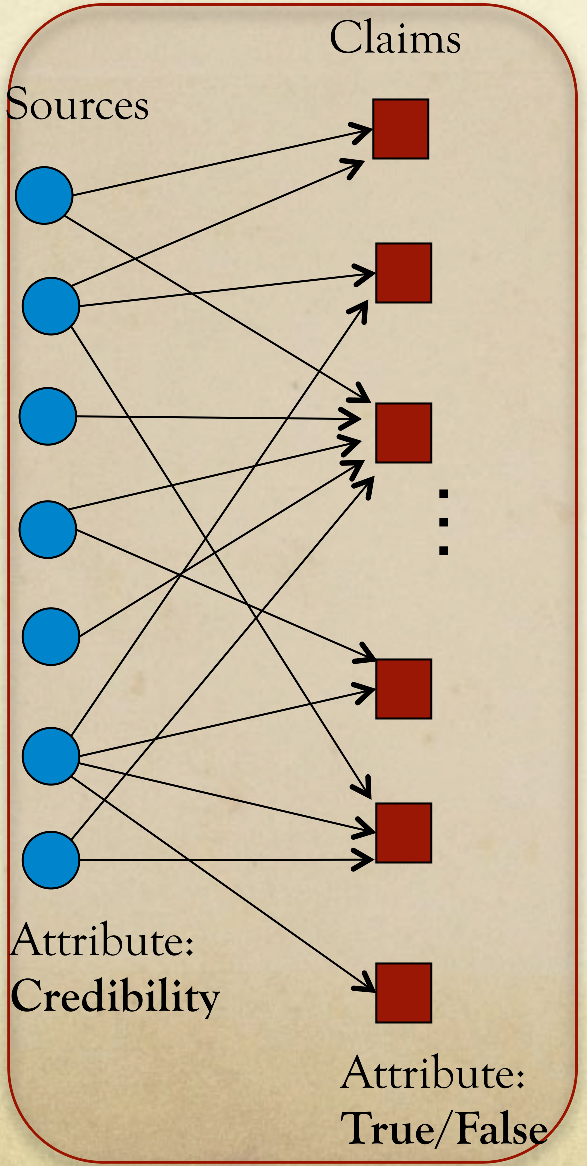
# Example Engine: Apollo

Events

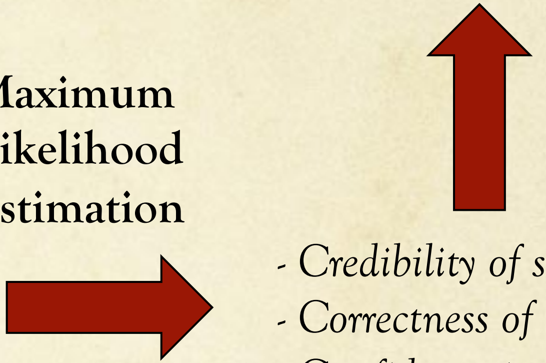
Egypt Unrest

Hurricane Irene

Fukushima



**Maximum Likelihood Estimation**



**Event Summary**

- *Credibility of sources*
- *Correctness of claims*
- *Confidence intervals*

- Formulate the fact-finding problem as one of maximum likelihood estimation
- Solve it using the *Expectation Maximization* (EM) algorithm
- Compute a bound on estimation accuracy (using the Cramer Rao Bound)

# Fact Finding

Events

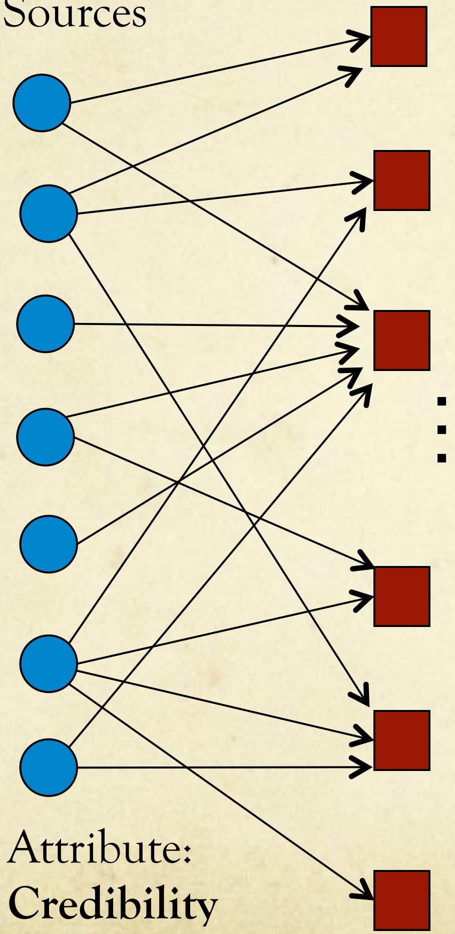
Egypt Unrest

Hurricane Irene

Fukushima

Sources

Claims



Attribute: **Credibility**

Attribute: **True/False**

**Maximum Likelihood Estimation**

**Event Summary**

- *Credibility of sources*
- *Correctness of claims*
- *Confidence intervals*

- Formulate the fact-finding problem as one of maximum likelihood estimation
- Solve it using the *Expectation Maximization* (EM) algorithm
- Compute a bound on estimation accuracy (using the Cramer Rao Bound)

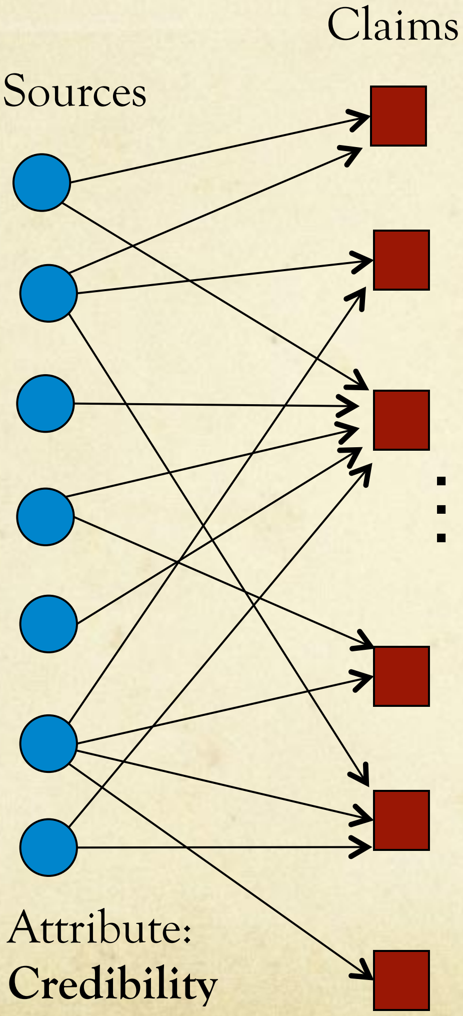
# Fact Finding

Events

Egypt Unrest

Hurricane Irene

Fukushima



Attribute:
**True/False**

Problem?

**Maximum Likelihood Estimation**

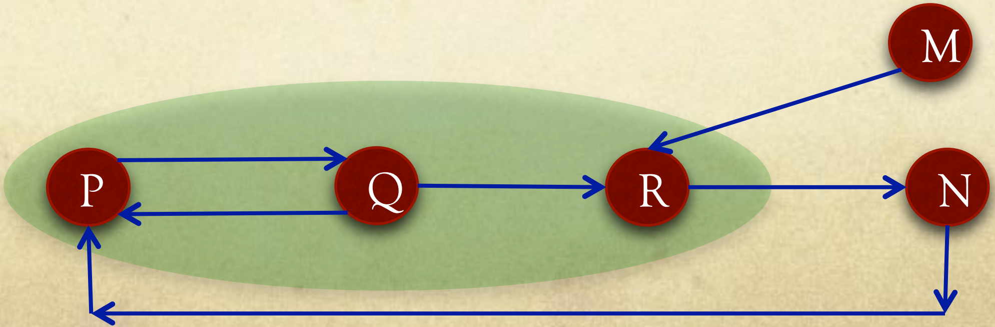
- *Credibility of sources*
- *Correctness of claims*
- *Confidence intervals*

**Event Summary**

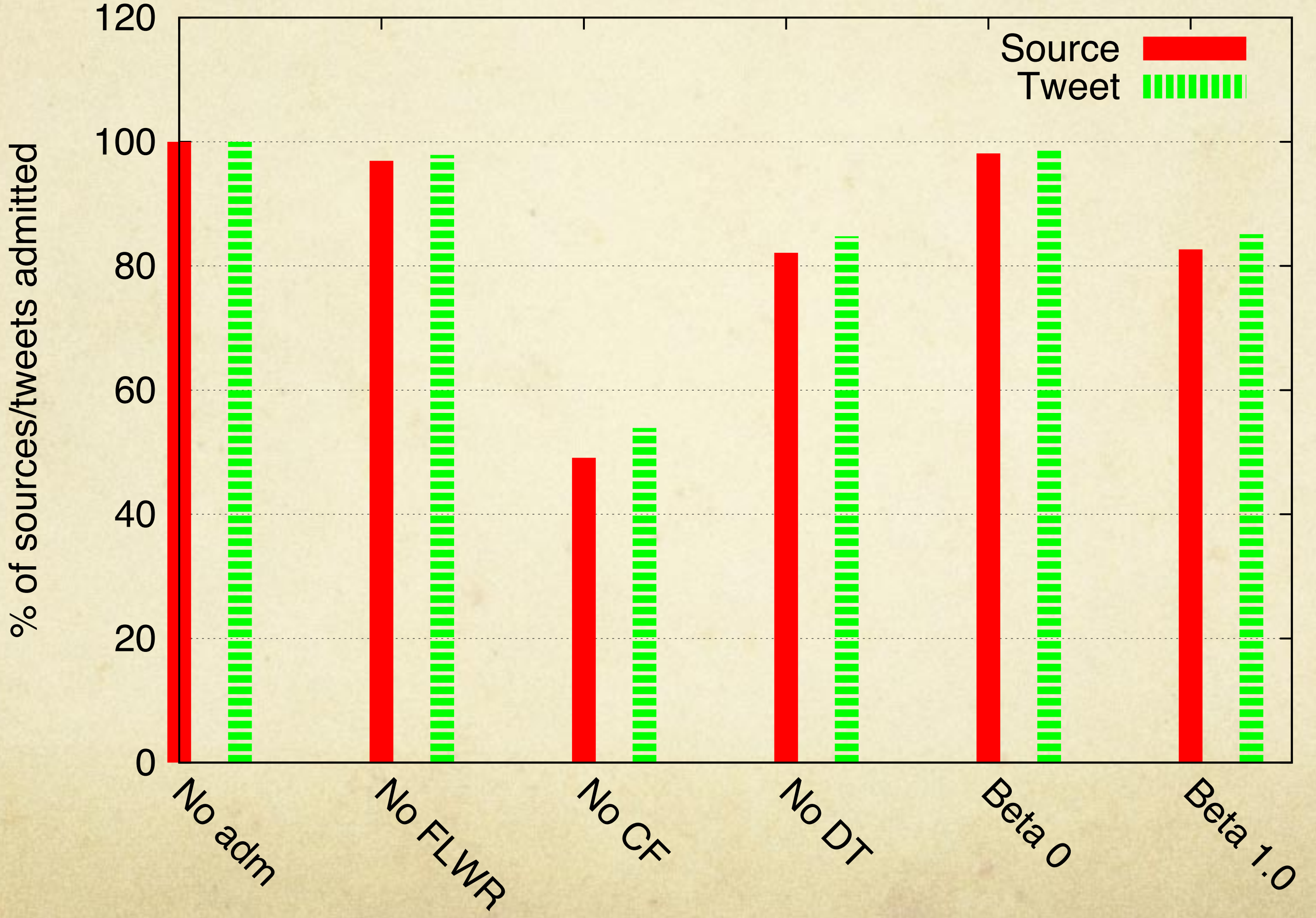
- Formulate the fact-finding problem as one of maximum likelihood estimation
- Solve it using the *Expectation Maximization* (EM) algorithm
- Compute a bound on estimation accuracy (using the Cramer Rao Bound)

# Admission Control Schemes

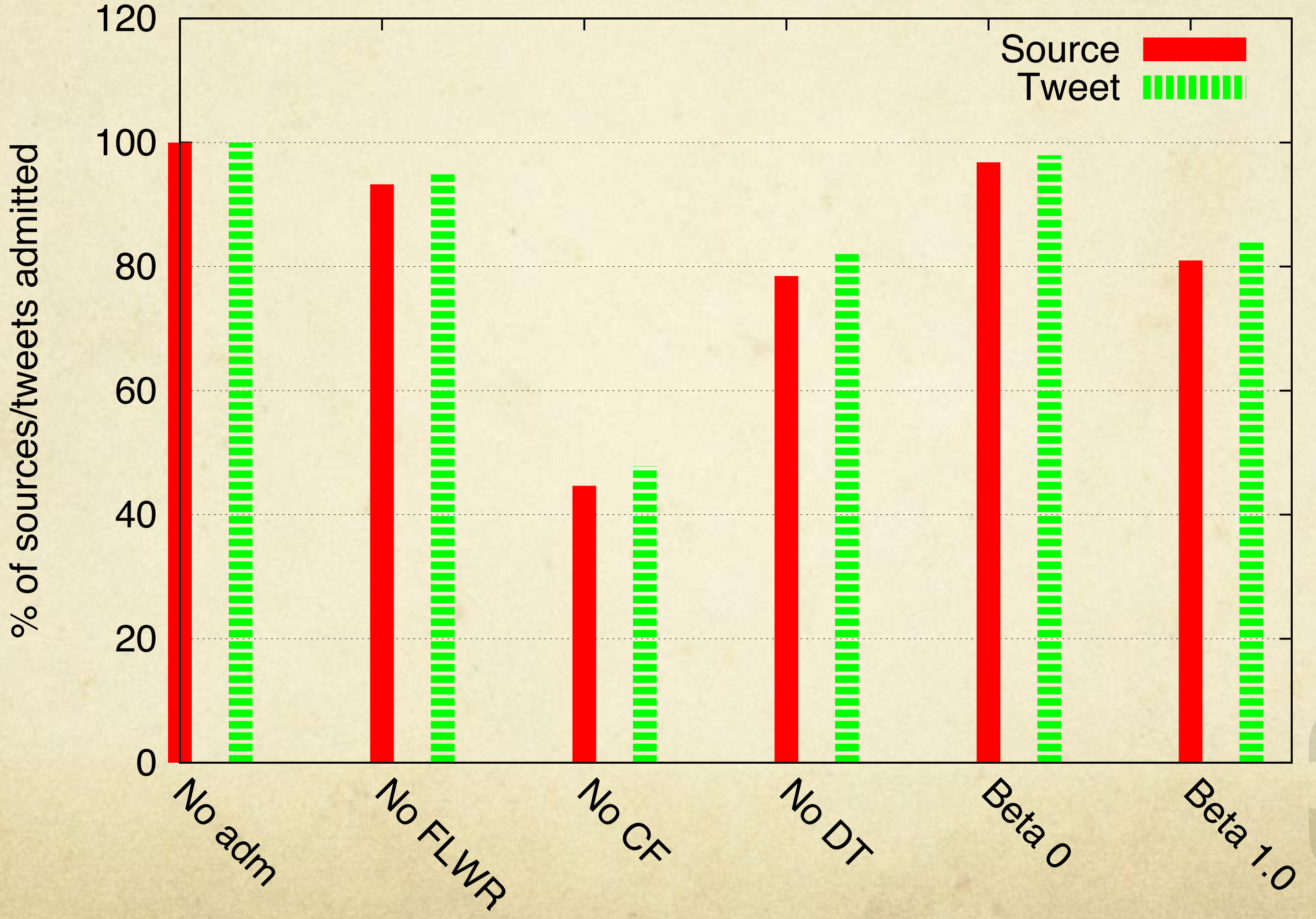
- $\beta$ - Controller : At each step, select the source if it progressively improves *Independence Score* of the set S, and its own independence Score exceeds $\tau$
  - Dependency function $f_{ij}$ taken to be $p^k$ , where $k$ is the length of path from $i$ to $j$. $p$ is taken to be 0.5
  - Therefore, if S = {P, Q, R} in the following graph,
    - $f_{PQ} = 0.5$, $f_{PR} = 0.25$, $f_{PN} = 0$, $\beta\,(P,S) = 0.5 * 0.75 = 0.375$
    - $f_{QP} = 0.5$, $f_{QR} = 0.5$, $f_{QN} = 0$, $\beta\,(Q,S) = 0.5 * 0.5 = 0.25$
    - $f_{RP} = 0.25$, $f_{RQ} = 0.125$, $f_{RN} = 0$, $\beta\,(R,S) = 0.75 * 0.875 = 0.65625$
    - $f_{MP} = p^3$, $f_{MQ} = p^4$, $f_{MR} = p$, $\beta\,(M,S) = p^3 + p^4 + p = 0.6875$
    - $B(S) = 0.375 + 0.25 + 0.65625 = 1.28125$



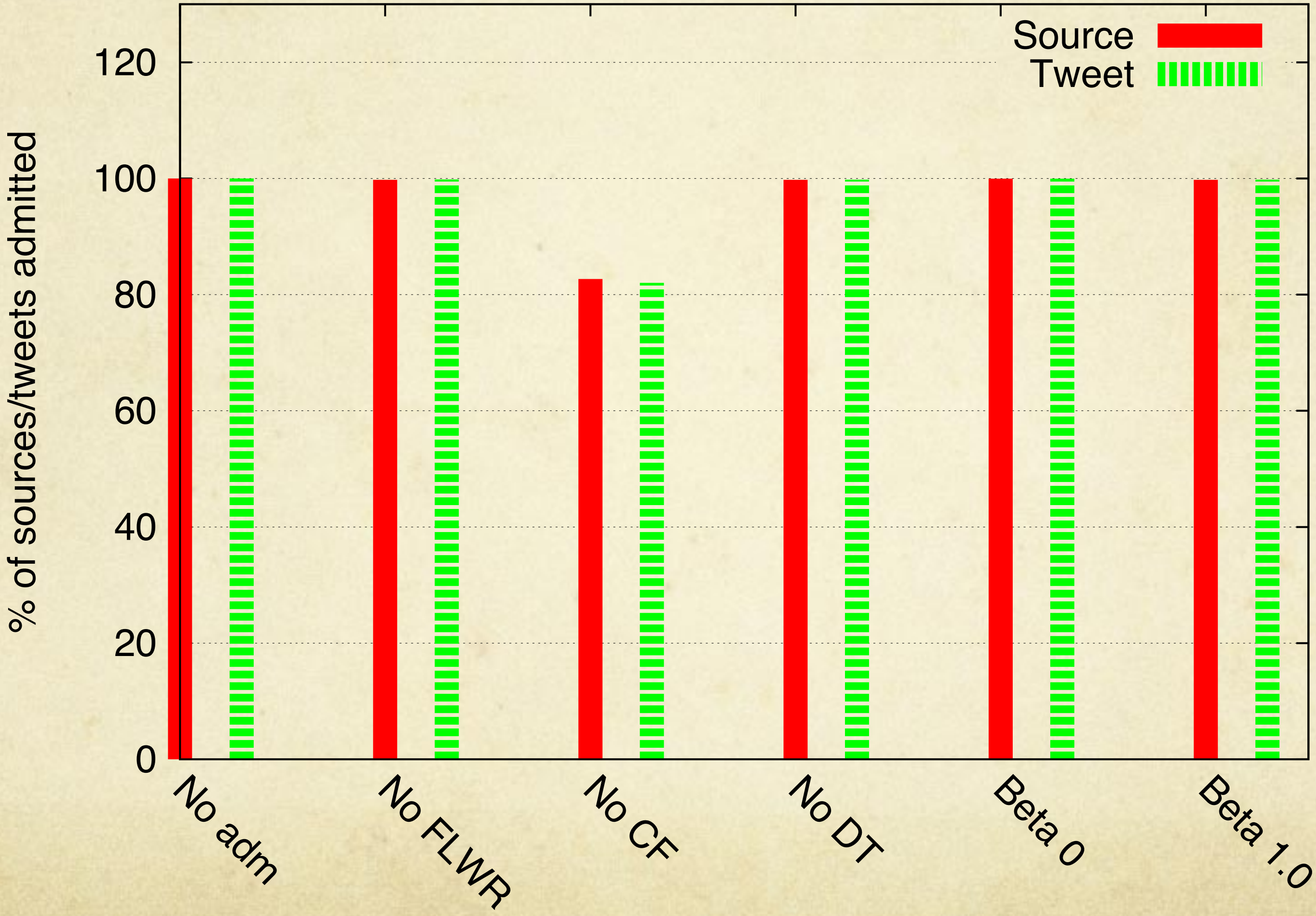
# Admission Statistics – Egypt (no RT)



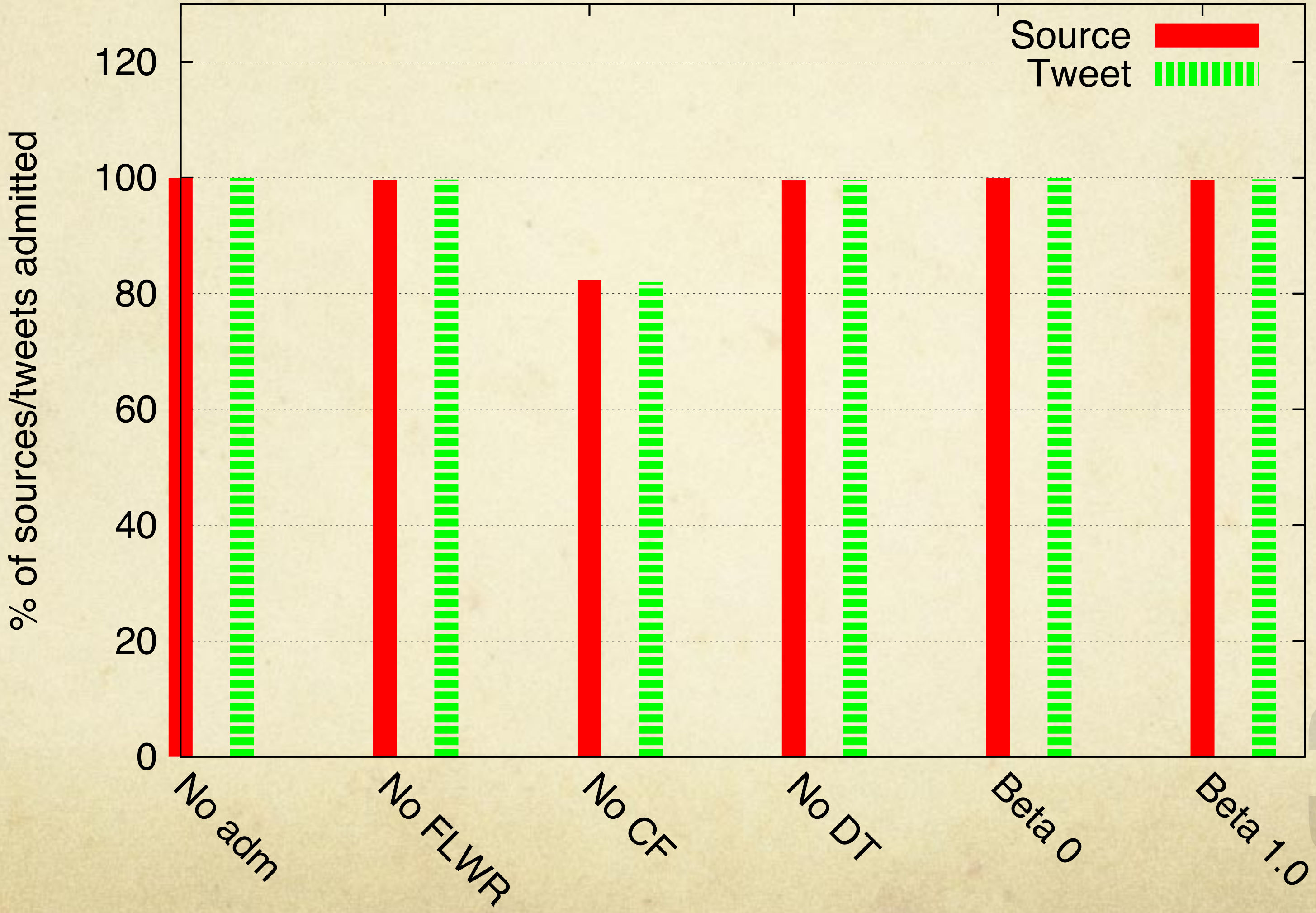
# Admission Statistics – Egypt (with RT)



# Admission Statistics – Irene (no RT)



# Admission Statistics – Irene (with RT)


Source
Tweet
% of sources/tweets admitted
No adm
No FLWR
No CF
No DT
Beta 0
Beta 1.0