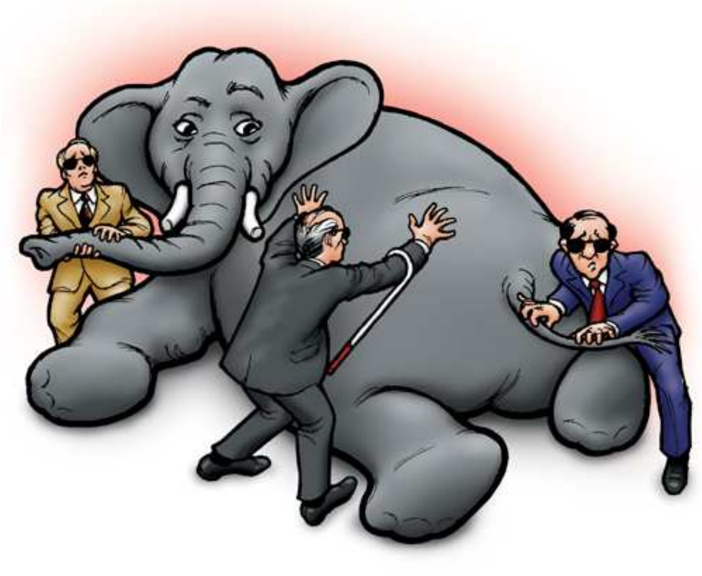
Learning and Ranking in Graph Data Models

Soumen Chakrabarti
IIT Bombay
http://soumen.in/doc/NetRank/

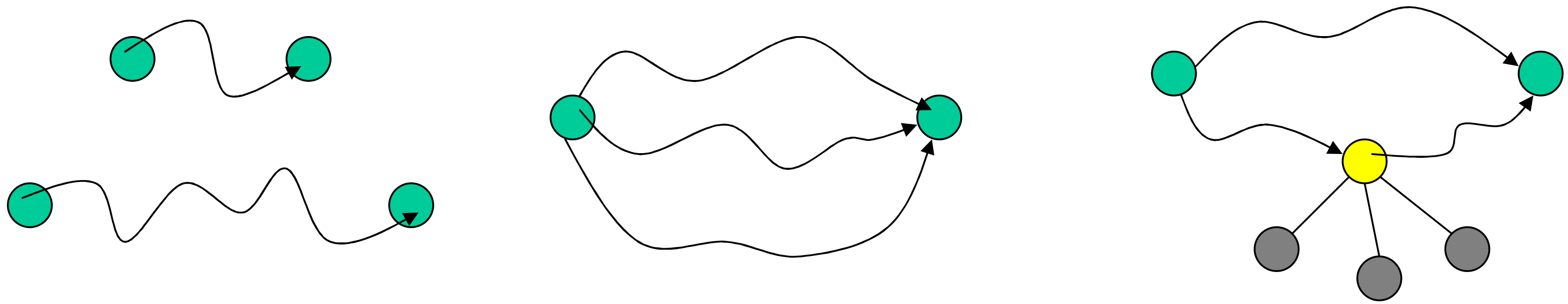
Abstract graphs

- Nodes, (binary) edges
- Edge weights, perhaps node weights
- Only one “kind” of node and edge
- Limited ability to represent real-world data
- Can already pose a number of difficult problems



Asymmetric influence

- How strongly does node u influence node v?
  - Length of path/s from u to v
  - Number of (edge disjoint) paths from u to v
  - Distractions on the way



- Random walks and electrical networks
  - Hitting time
  - Effective conductance

Symmetric similarity

- How similar are nodes u and v?
- How similar are their neighborhoods?
- Let $N(u)$ be (in/out/both) neighbors of u
- Base case: $s(u,u) = 1$
- PageRank on squared graph

$$s(u,v) = \frac{\alpha}{|N(u)||N(v)|} \sum_{p \in N(u), q \in N(v)} s(p,q)$$

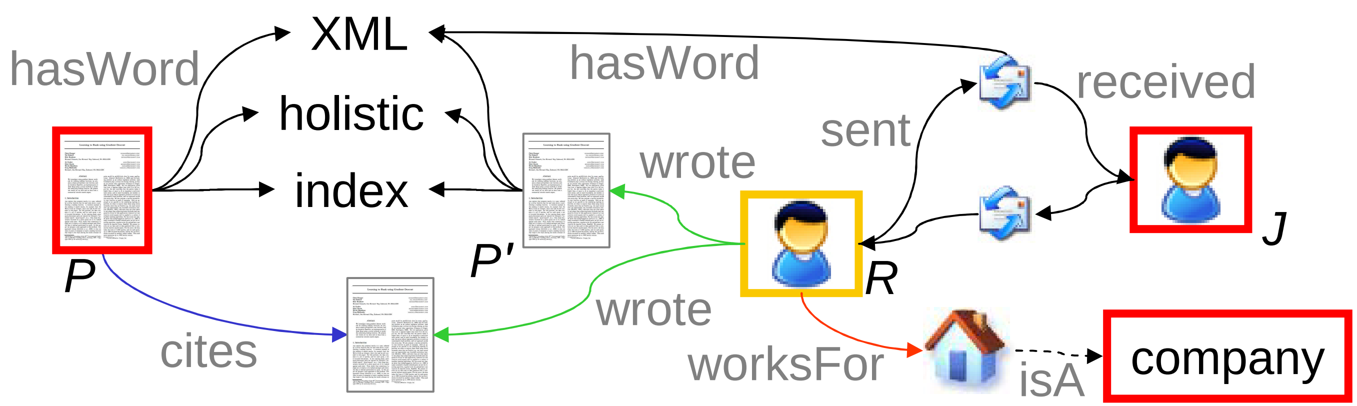
Missing link: Low-rank factors

- Should there be an edge between u and v?
- Not necessarily (just) because they are (directly) similar
- Adjacency matrix A is noise added to a low rank matrix UV
- Edge weights +1, –1, 0 (don't know)

$$\min_{U,V} \sum_{i,j} |A_{ij}| \max\{0, 1 - A_{ij}(UV)_{ij}\}$$

Real-world complications

- Nodes
  - Have types: person, organization, email
  - Are associated with feature vectors: dob, pan
- Edges
  - Have types: worksFor, wrote
  - Are associated with feature vectors: emailDate
- Hyperedges (for general relational data



Node labeling/scoring/ranking

- *The* problem in graphical models
- In general, hard; easier special cases
- Smoother models for associative potentials
  - Edge $\{i,j\}$ has association strength A_{ij}
- Node i associated with feature vector x_i
- Local score $s_i = w \cdot x_i$, final score f_i

$$\min_{w,f} \sum_i (w \cdot x_i - f_i)^2 + C \sum_{\{i,j\} \in E} A_{ij}(f_i - f_j)^2$$

- Laplacian smoothing

"Inverse" PageRank

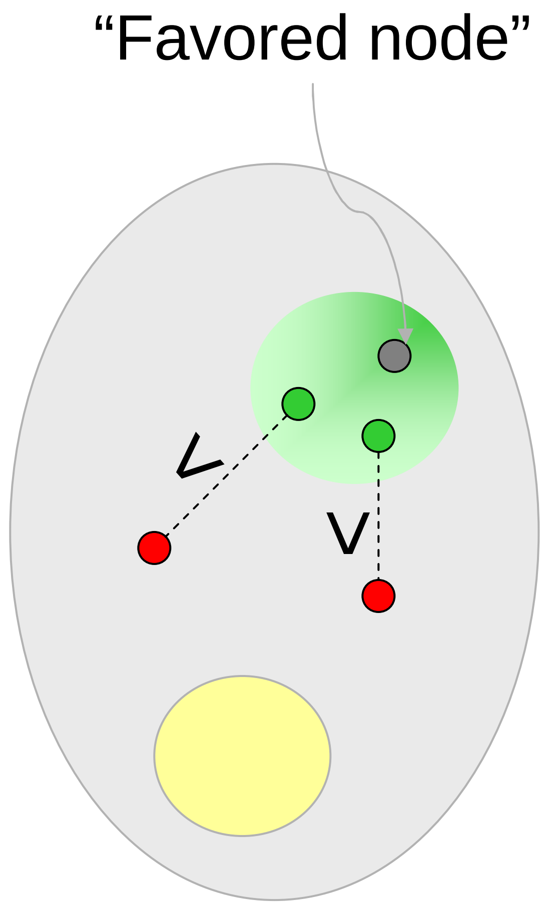
- Original PageRank: edge conductances fixed, find influence (effective conductance) of (from) one node on (to) all others
  - And rank them by decreasing influence
- Inverse PageRank
  - Given graph skeleton but not edge conductances
  - Given sampled partial comparison between pairs of nodes wrt influence
  - Infer edge conductances
  - So as to generalize influence to other nodes

Preferred community scenario

- Ranking papers for Data Mining researcher
- Some subgraphs and citations more important than others
- Revealed via pairwise preferences
- Do not estimate $C(j,i)$ directly
- Directly estimate p_{ij}, a constrained "flow" from i to j
- "BTW" $C(j,i) = p_{ij} \Big/ \sum_{(k,i)\in E} p_{ki}$

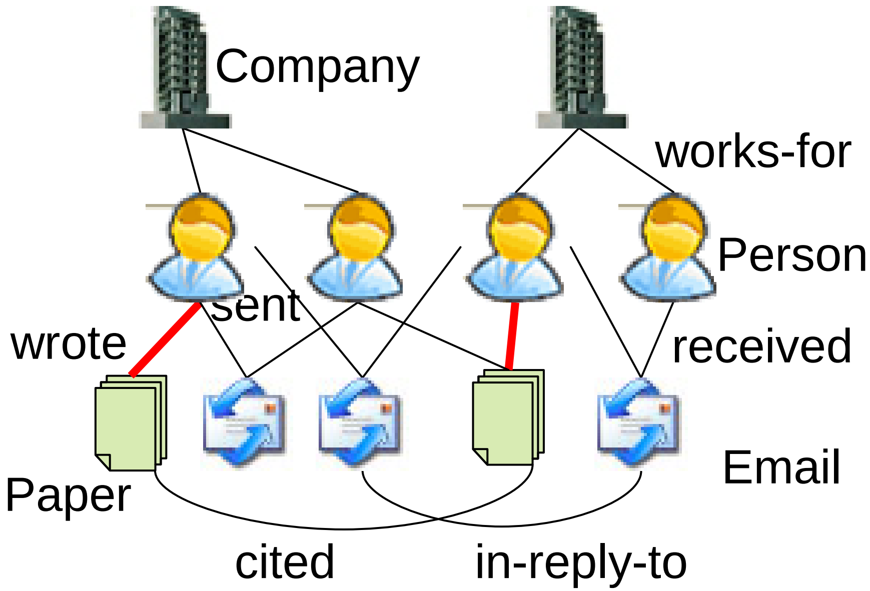
Inflow into i

- Local "transductive" setting
- Lots of parameters



Entity-relationship graph scenario

- Many node and edge types
- Edge e has type $t(e) \in \{1, \ldots T\}$
- Weight $w(i,j) = \beta(t(i,j))$
- Find $\beta(1)$, $\beta(2)$, …, $\beta(T)$ for least violation
- "Global entanglement" but far fewer parameters
- Somewhat "inductive", can augment graph with objects of known types

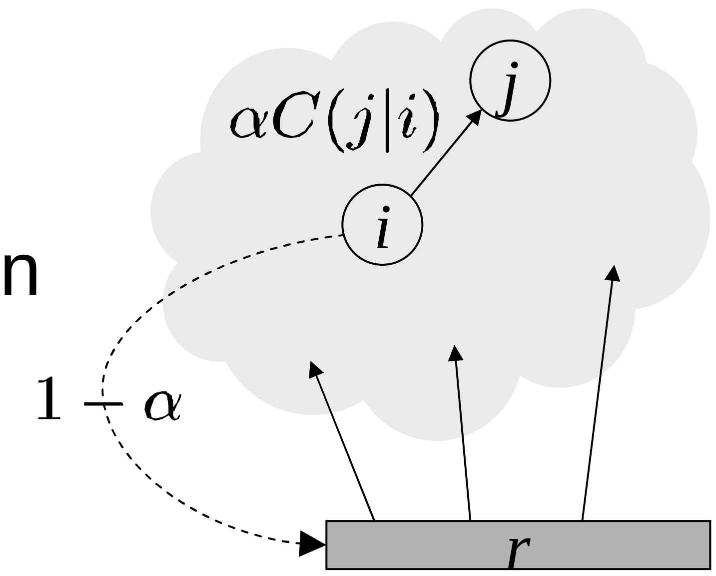


PageRank: Conventional view

- Inputs
  - Graph with edge conductance matrix *C*
  - Personalized teleport distribution *r*
  - Walk with probability α, teleport w.p. $1-\alpha$
  - "Biased random surfer"

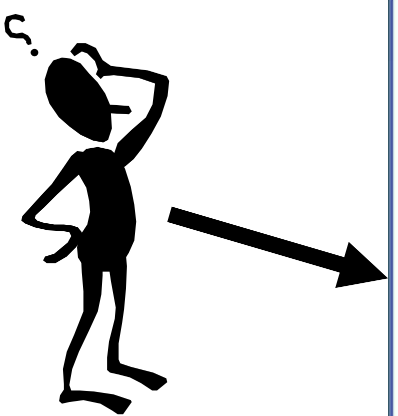
$$p = \alpha C p + (1-\alpha) r$$

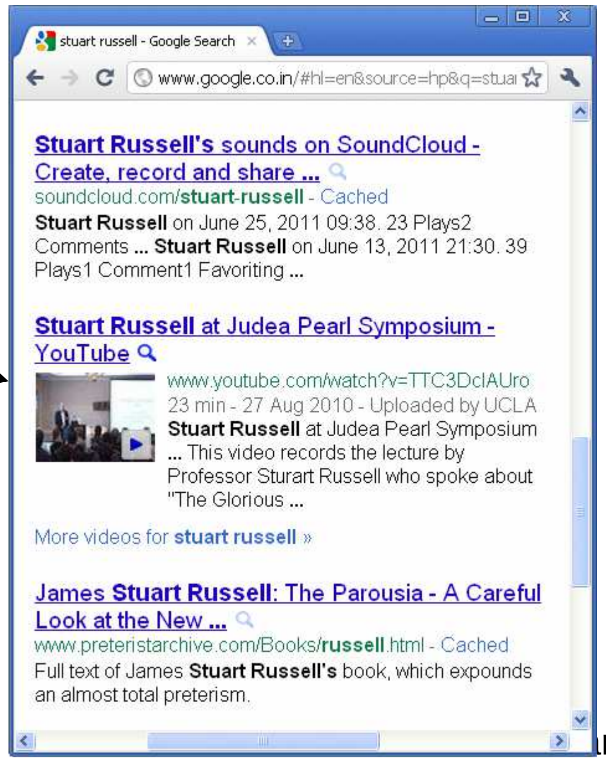
- Output
  - Steady state visit distribution
  - "You should emulate the aggregate behavior of many random surfers"

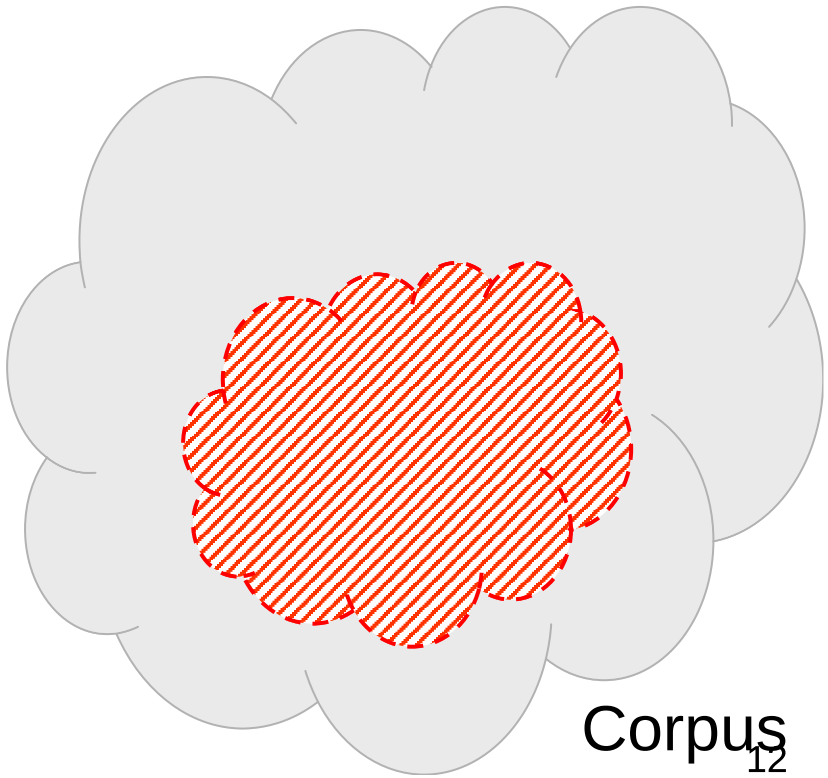


User view: Exact opposite!

- ~~Random~~ **search-guided** surfer
- Search engine knows **relevant subgraph**
- But user can inspect only a few hits
- Search engine **outputs sparse** teleport *r*

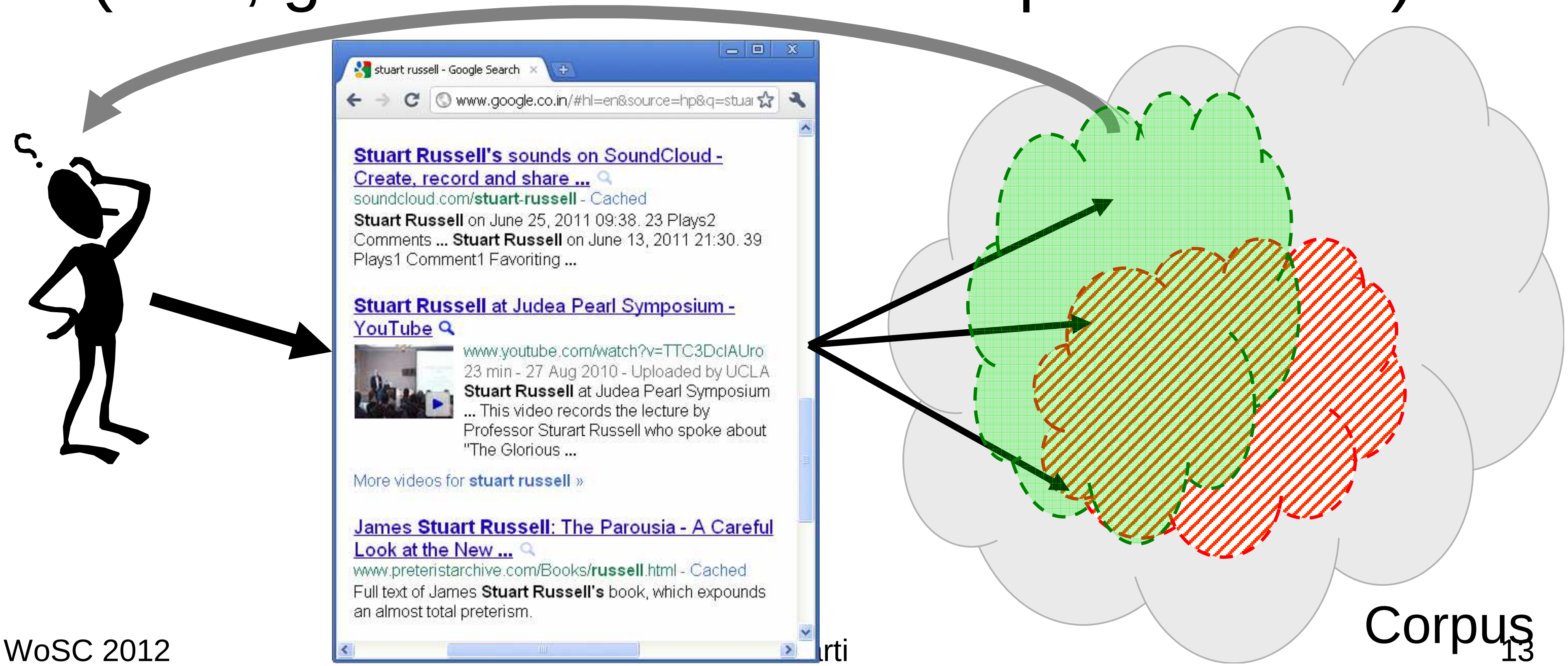






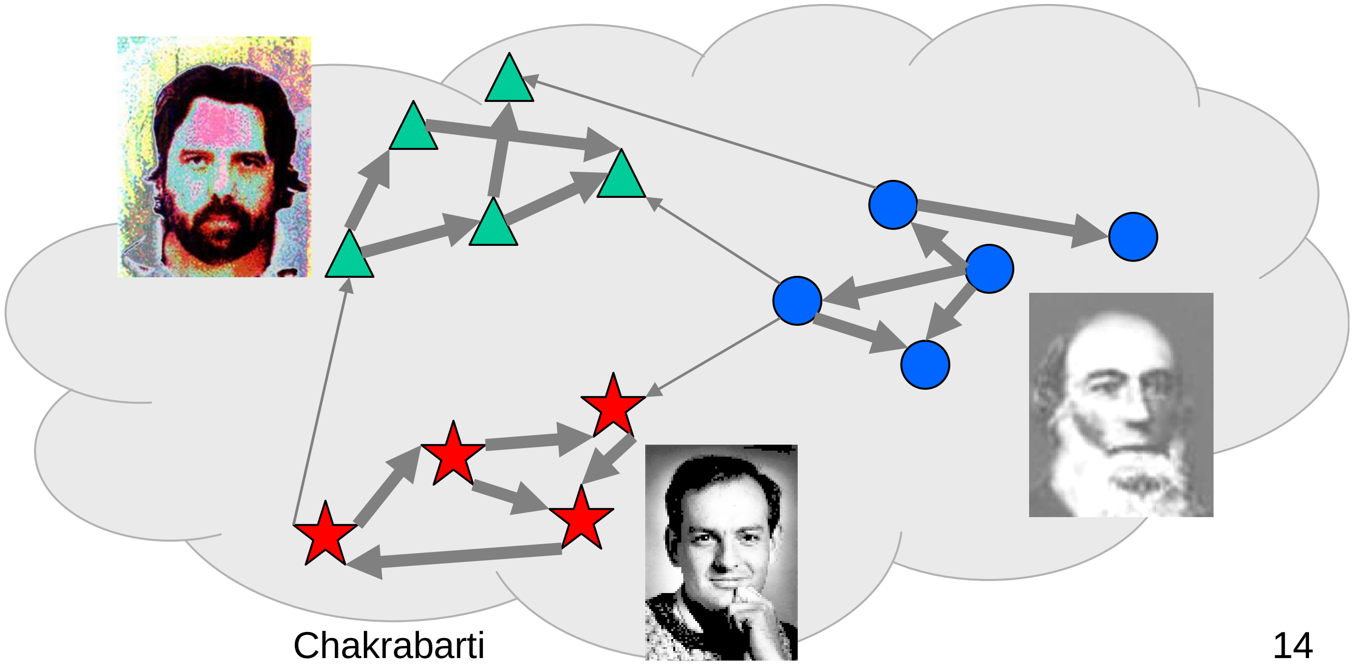
User view: Exact opposite!

- User **diffuses** out through sparse teleport
- Occasionally teleports back to search results
- Eventually explores **green** subgraph
- (Red, green “boundaries” are probabilistic)



Diffusion defined via subsumption

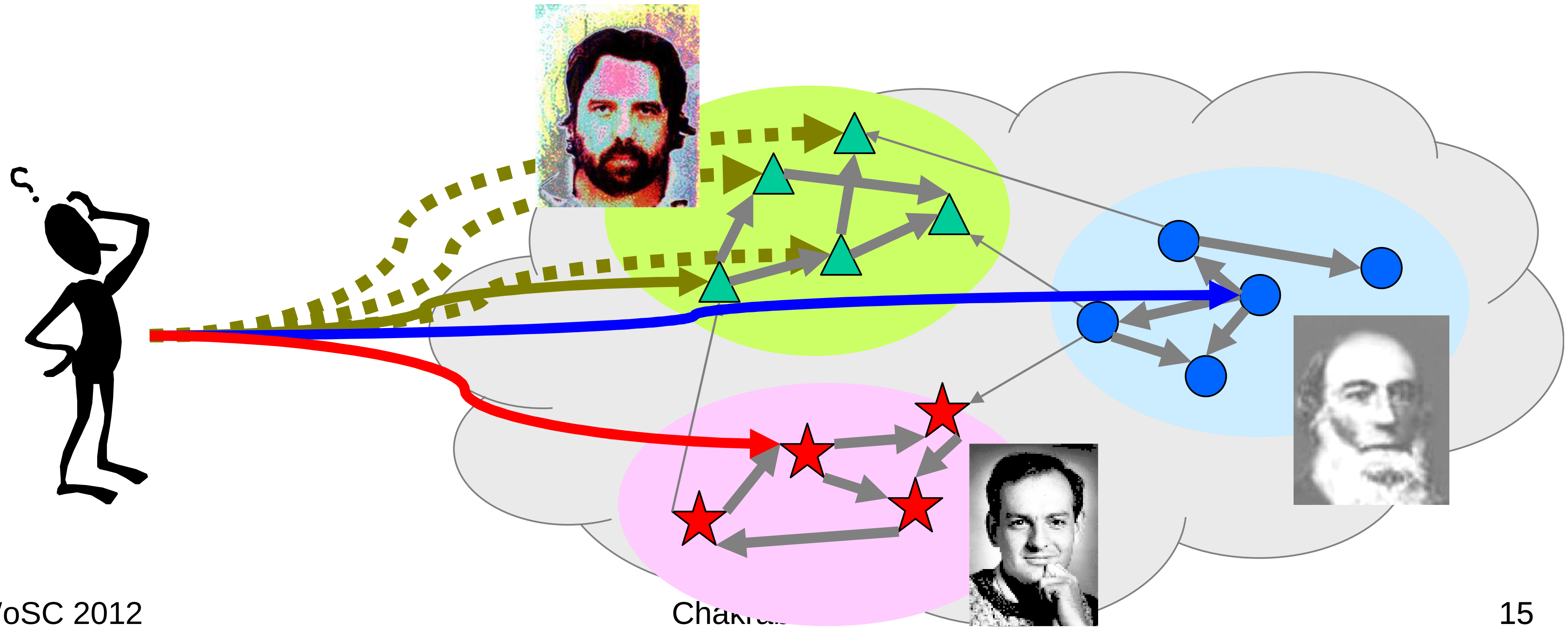
- Original PageRank: diffusion via hyperlinks
- But frequently used with other kinds of edges
- Suppose surfer is on page *i*
- And, having read *i*, there is no new info in *j*
- Then let $C(j|i)$, also written as $C(i \rightarrow j)$ be **large**



Graph center diversity (GCD)

- Suppose the searcher can click through at most three links returned by the search engine
- If any of the pages could be potentially relevant, …
- … then we cannot waste teleports on one cluster

A natural definition of diversity



Formulation summary

- Search engine knows what's best for query
  - Node *i* has relevance *b*(*i*)
- User has limited patience scanning results
  - *r* must be sparse: at most *K* positive elements
- Conductance matrix C and walk probability α predict user behavior once given *r*
- Steady state visit probabilities given by

$$(1-\alpha)(\mathbb{I}-\alpha C)^{-1}r$$

- **Inference, hard:** design sparse *r* to minimize

$$\left\|\vec{b}-(1-\alpha)(\mathbb{I}-\alpha C)^{-1}r\right\|$$

Infection origin problem

- Observe node “infection” for a while
- But starting some time after the infection was first introduced
- Trace back (probabilistically) to the origin node(s)
- Obviously, impossible to reduce entropy on a complete graph
- What graphs are amenable to such forensics?
- Do the infected ever get immune/cured?

Marketing problem

- EvilCorp wants all kids to eat sugary candies
  - Or their dads to buy iPhones
- Obtain social network with edge strengths
- Have finite marketing budget
- Celebrities expensive to convert, but they influence a lot of people
- Allocate finite budget most judiciously

Concluding remarks

- Graph have always been a (too?) powerful data model
- Many formulations and approaches for mining “abstract” graphs
- Real-world data turns into graphs with additional info (node, edge features, time)
- More work to do on learning and ranking problems on real-world graphs