Social Data

Daniel Huttenlocher

Dean of Computing and Information Science
Cornell University

Scientific Potential of Social Data

- Online social interactions yielding fundamentally new sources of data
  - Direct record rather than recollection or observation
- Means for making sense of this data currently in early stages
  - Challenge for both social sciences and for computing and information sciences
- Online behavior ahead of our understanding
  - Broad but poorly understood implications for society: health, education, commerce, leisure, …

Computational Social Science

- Social sciences: studies and models of human interaction
  - Social psychology, sociology, ethnographic methods
  - Generally small but often rich settings
- Computing and information sciences: data mining and machine learning
  - Evaluating models with large datasets
  - Often fairly “impoverished” models
- Bridging gap of different mindsets and methodologies

Today's Talk

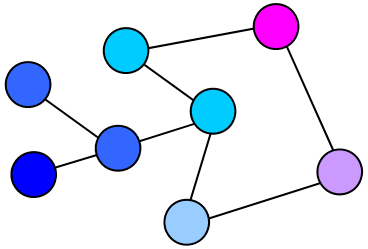
- Some of our recent work at Cornell, bringing together classical theories and computational tools
  - Interdisciplinary group in Comp Sci, Comm, Econ, Info Sci, Mgmt, Psych, STS, Soc
    - But talk covers CS-centric material
- Part I – similarity and influence in social networks: theories of homophily
- Part II – signed networks: theories of friendship and status
- Part III – shared perception of places

Part I: Similarity in Social Networks

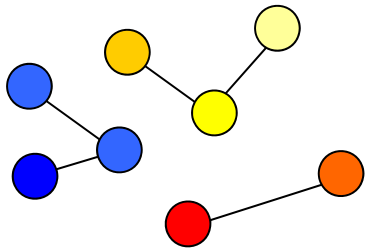
- Homophily: tendency to be similar to neighbors in social network
  - "Birds of a feather"
  - Refers to both outcome and process
- Two underlying processes
  - *Social influence* – people adopt behaviors of those they interact with
    - Interactions influence interests
  - *Selection* – people tend to form relations with those similar to them
    - Interests influence interactions

Roles of the Two Processes

- Both result in neighbors looking similar, possibly different structural characteristics
  - Social influence encourages homogeneity whereas selection encourages fragmentation



Homogeneity



Fragmentation

- Viral marketing depends on social influence whereas recommender systems rely only on similarity (from influence or selection)

Consider Two Basic Questions

- Can we quantify and model the interplay between social influence and selection?
  - Better understand how these two phenomena interact in the creation of social networks
  - Model relation between similarity and social interaction
- To what extent do similarity and social interaction affect future behavior?
  - Predictors of what people will do, relative to one another
- On social media and social networking sites

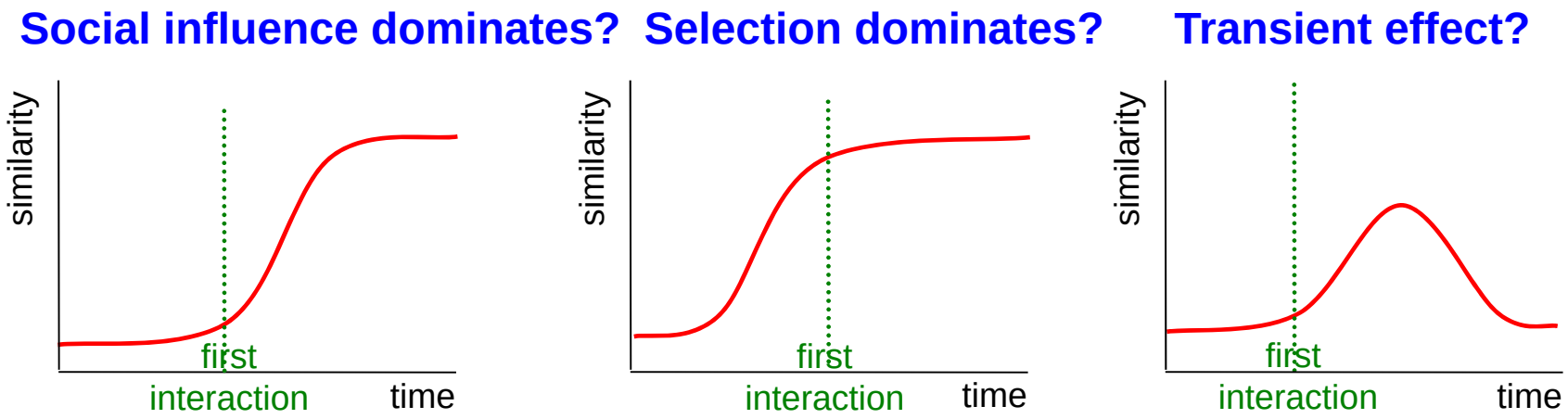
Social Network of Wikipedia Authors

- Task-focused activity of writing articles, interaction through discussion (talk) pages
  - Consensus and disagreements
- Article edits as activities
  - 53M edits of 3.4M articles
- User talk edits as social interactions
  - 510K users (doing 61% of article edits)

English Wikipedia, 2007 dump

Interplay of Influence and Selection

- Both social influence and selection suggest similarity should increase with interaction

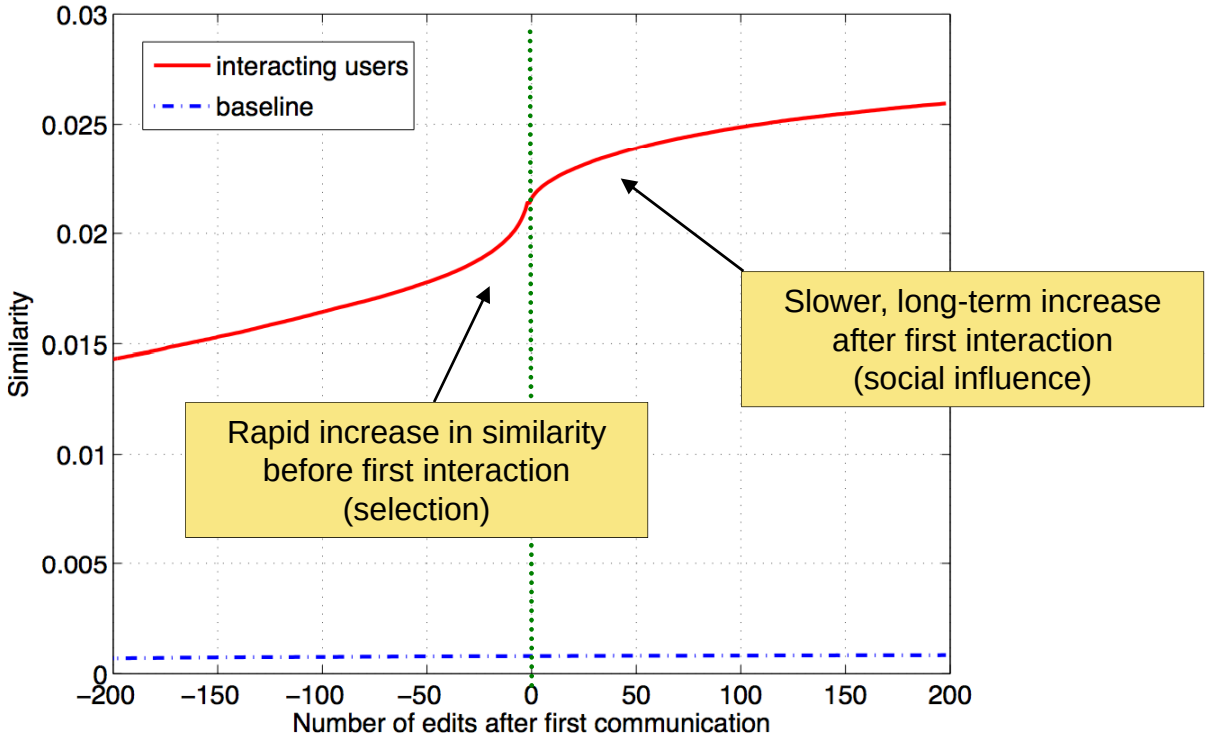


- Aggregate similarity of pairs of users over time, aligned by first interaction

Measuring Similarity

- History of activity captures user interests
  - In Wikipedia, editing articles
  - Generally varies over time
  - Vector space model: at time *t*, each user has an *m*-vector giving degree of involvement in each of *m* articles, weighted or binary
- Compare history vectors at given time to quantify similarity of interests
  - Many ways of comparing these vectors, does not effect overall results
    - Cosine, Jaccard distance, tf-idf, etc.

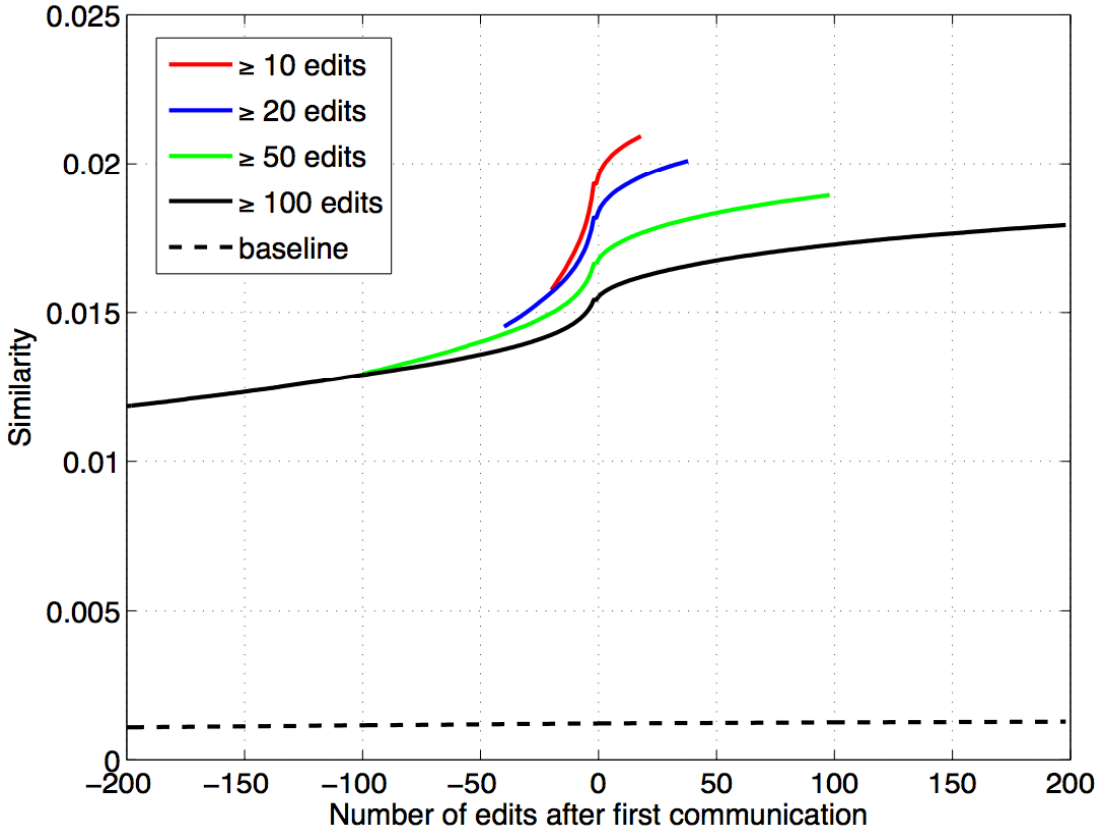
Wikipedia Similarity vs. Interaction



- All pairs of users who interacted vs. baseline for pairs who never interacted aligned at arbitrary time
- Cosine measure (others similar)

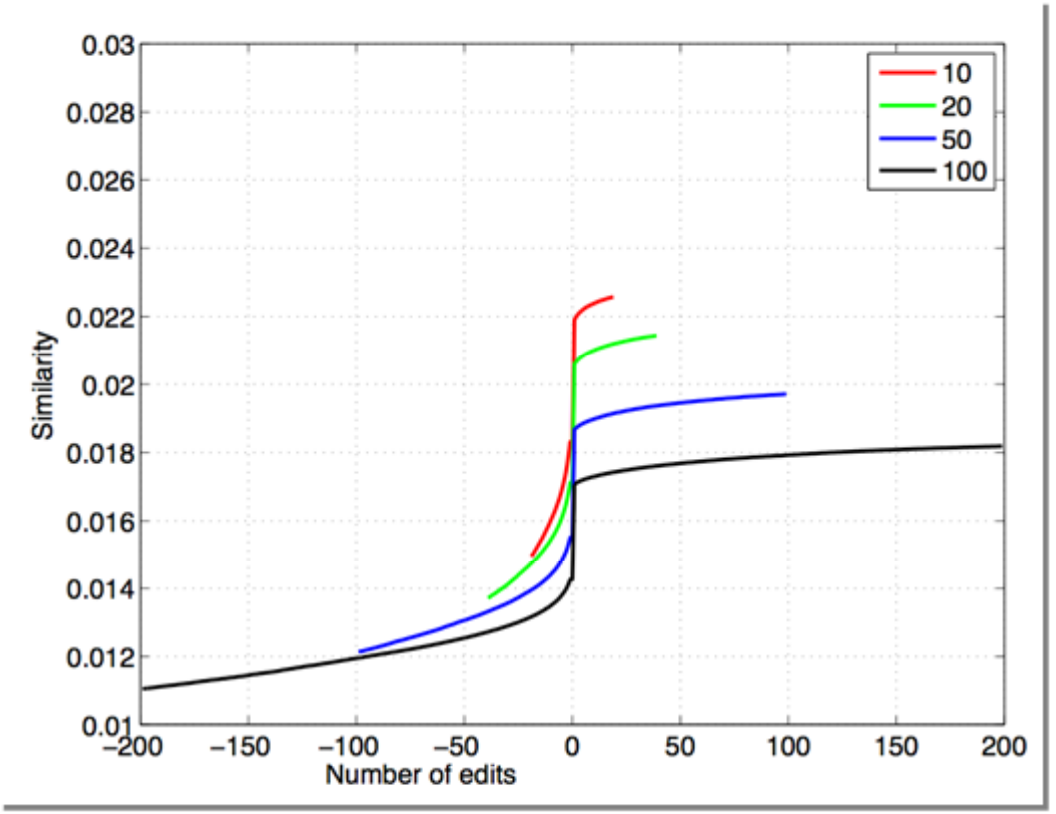
Wikipedia Similarity vs. Interaction

- Stable across users with different activity levels



Wikipedia Similarity vs. Interaction

- Effects present before first and after last communication

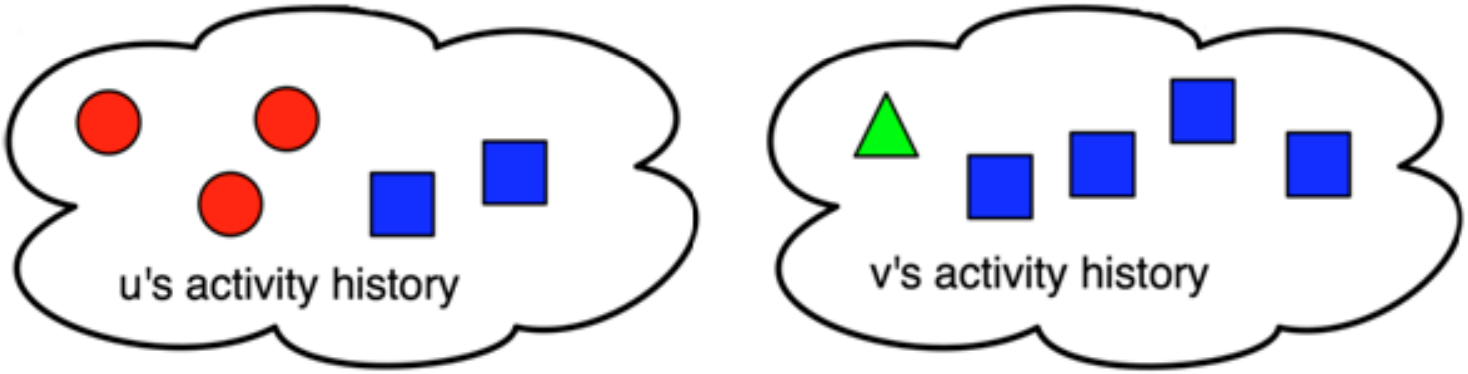


Findings on Similarity and Interaction

- Similarity between people increases rapidly before they first interact
  - E.g., random walks through interest space, chance of meeting depends on distance
  - Evidence for a selection process
    - Choose to interact with those of similar interests
    - More exposed to those share most with
- Similarity between people continues to increase long after they first interact
  - Evidence for long-term social influence not just short-term coordination

Modeling User Behavior

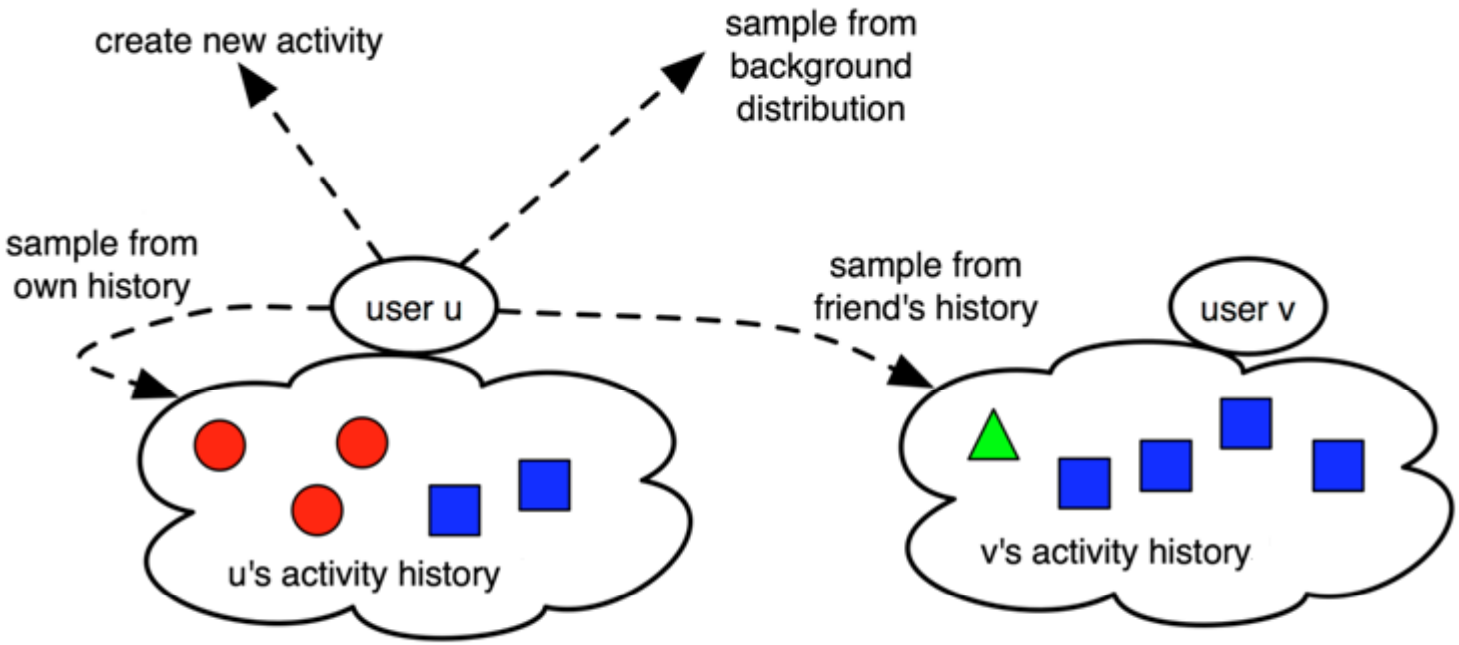
- Systems where people interact while undertaking activities
  - Each user has history of most recent activities



- Select activity from own previous activities or those of social ties
- Urn models, heavy tailed distributions

User Model

- At each time step, given user *u* either
  - Interacts with another user or performs activity
- Motivated by earlier model [Holme-Newman 06]
  - Hold single opinion



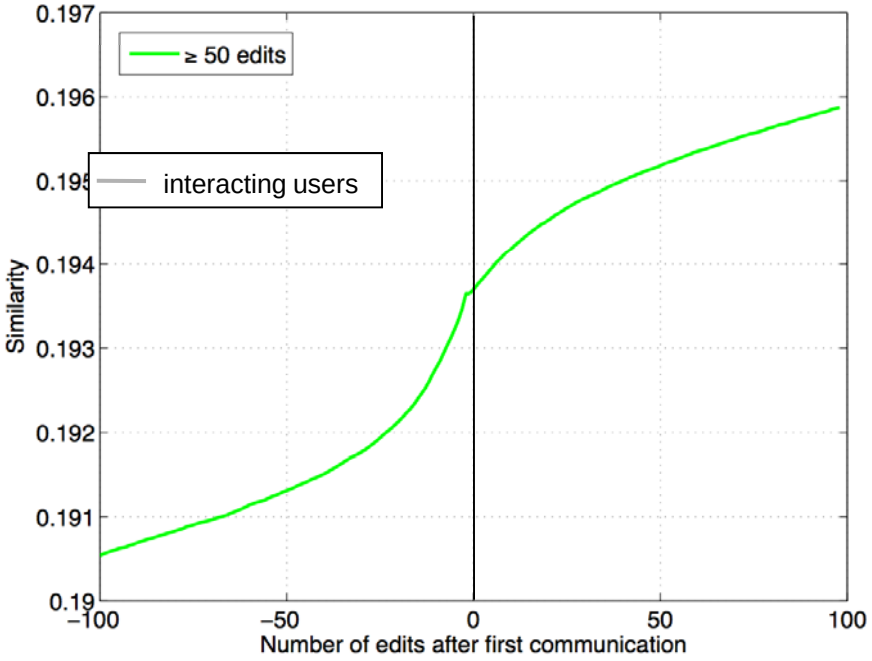
Wikipedia Parameter Values for Model

- Activities (article edits) greatly outnumber communication on user talk pages
  - About 16:1
- When communicate, do so with user with (previous) shared interests 29% of time
- When edit, do so
  - 35% of time based on own interests
  - 8% on neighbor's interests
  - 50% on "community" interests (exogenous)
  - 7% creating new article

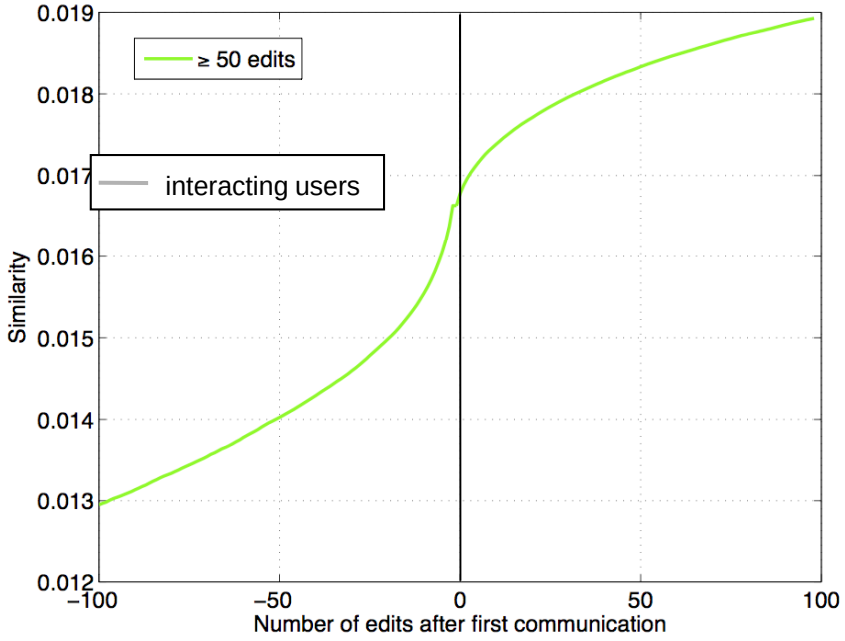
Simulation Qualitatively Similar

- Monte Carlo simulation of user behavior
  - 100K users, 3M timesteps

Simulated Wikipedia result



Actual Wikipedia result

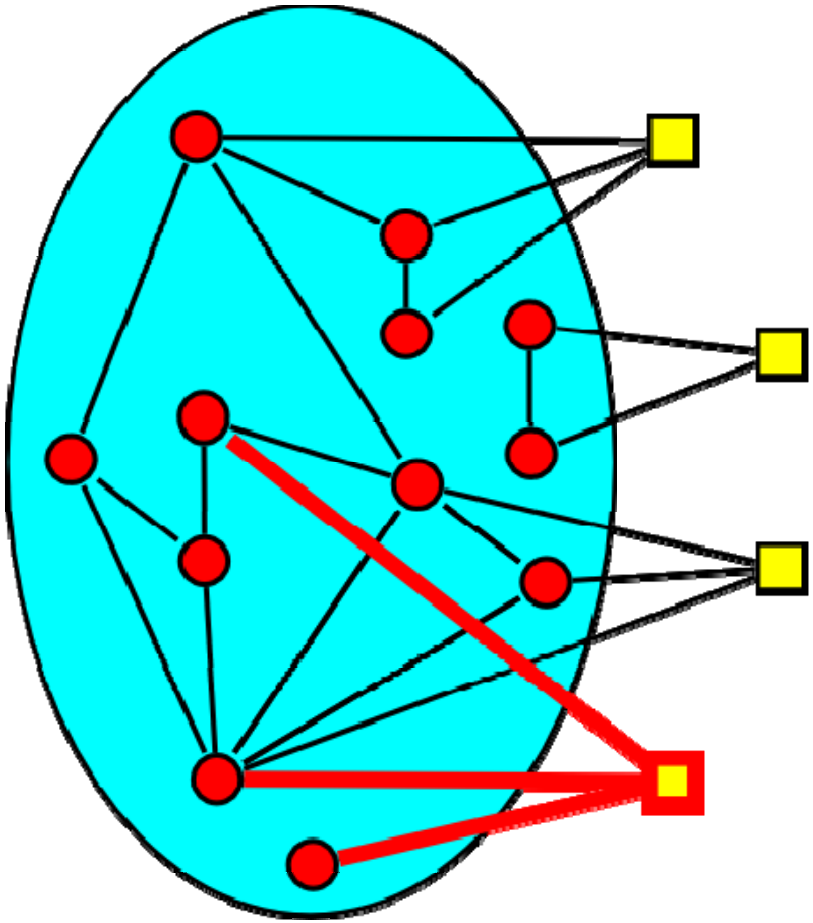


Reprise: Two Basic Issues

- Modeling the interplay between social influence and selection
  - Networked urn model produces qualitatively good fit to observed changes in similarity vs. number of interactions
- Extent to which similarity and social interaction affect future behavior
  - Degree to which are they predictors of what people will do, relative to one another
  - Propensity to engage in new activities as function of social ties – in aggregate

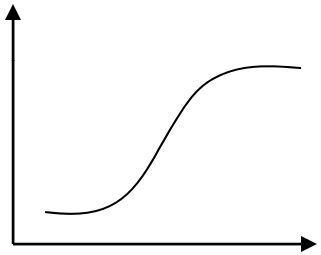
Engaging in New Activities

- As function of local network
- Red circles represent those in group, yellow squares might join
  - 3 ties vs. 2 ties
- Other structural features
  - E.g. how connected are ties to one another
    - Triads

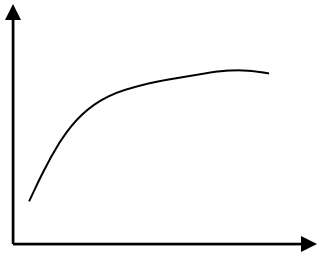


Probability of Engaging in New Activity

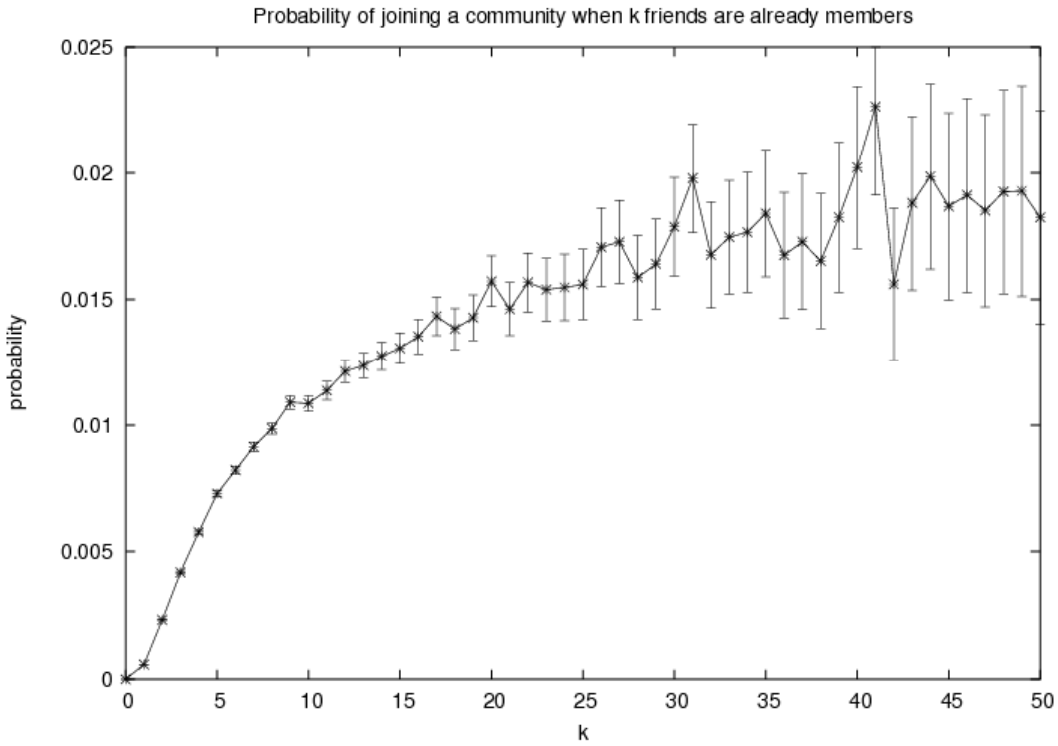
- As function of number of social ties – as opposed to adoption curves based on time
  - S-shaped? Critical mass effect (logistic)



  - Concave? Diminishing returns (logarithmic)

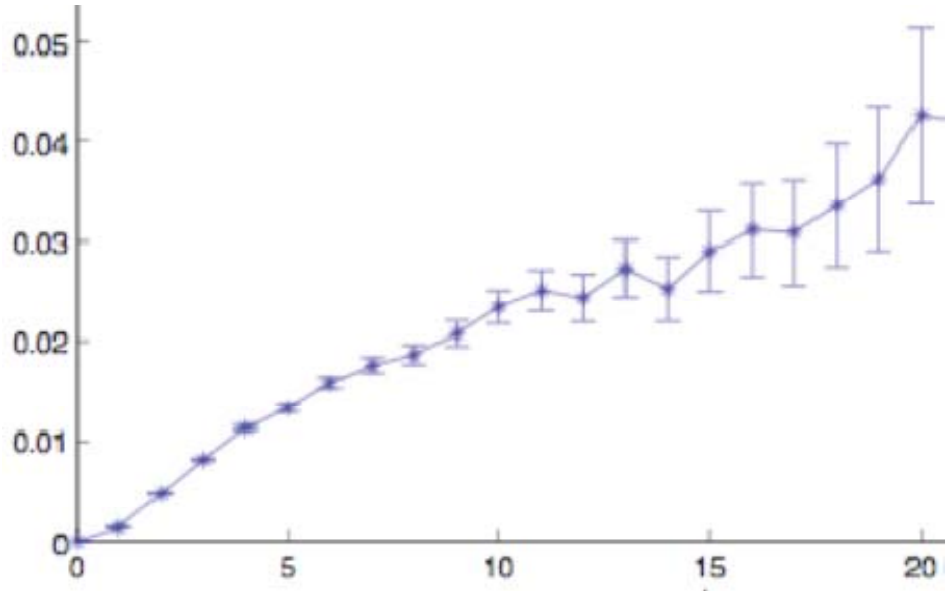


Joining LiveJournal Groups



- Supra-linear up to 2, then diminishing returns (even for large number of friends)
  - 12M users, 250K groups, over 2 month period

Becoming Editor of Wikipedia Page



- Diminishing returns, flatter overall shape
- Still supra-linear for 0-1-2
  - "Once is an accident twice is a pattern"

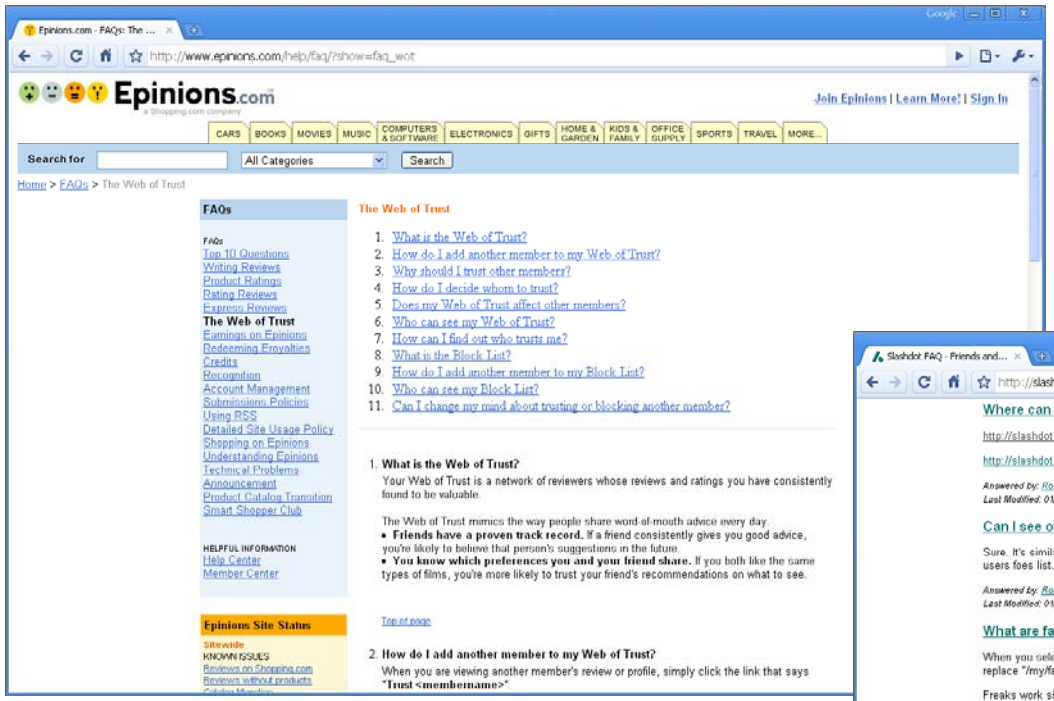
Part I Recap: Influence and Selection

- Evidence for selection
  - Increasing similarity before first interaction
- Evidence for social influence
  - Continued long-term increase in similarity after interacting
  - Probabilistic model indicating activities influenced by neighbors
  - Probability of engaging in new activity increases with number of neighbors already engaged in that activity
    - Diminishing returns except for 0-1-2

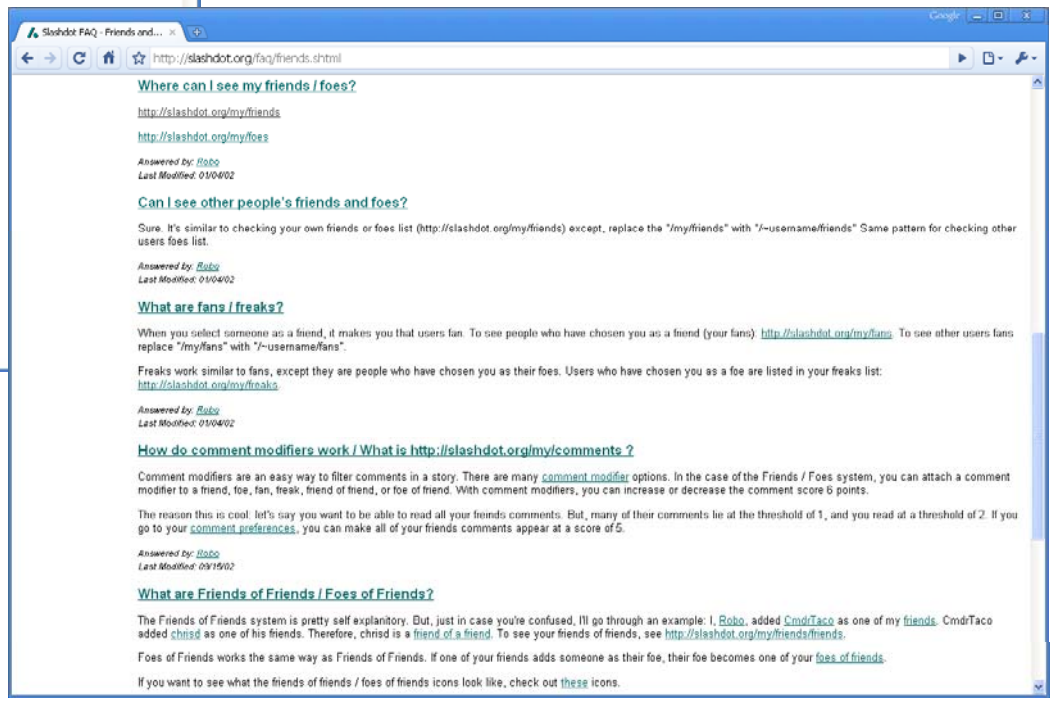
Part II: Signed Networks

- Mix of positive and negative relations between users on many social media sites
  - Yet most research on unsigned networks
- Investigate closed triads – relations between triples of people
  - In three online settings: Epinions, Slashdot, Wikipedia admin voting
- Compare classical and new theories
  - Structural balance from social psych [Heider 46]
  - Directed links as assessments of relative status (asymmetric in contrast with friendship or trust)

Explicit Signed Relationships



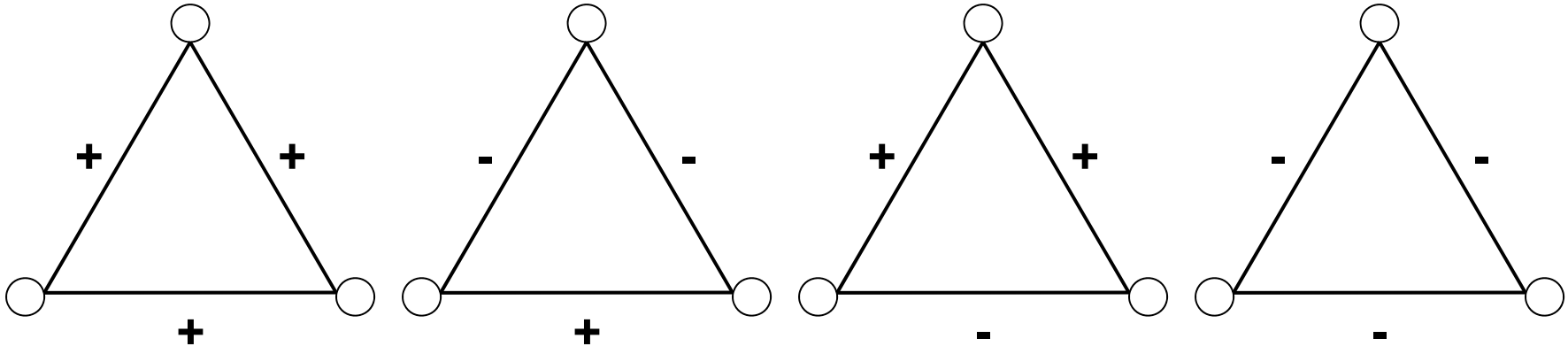
Epinions web of trust (and distrust) – distrust private



Slashdot friend or foe – publicly visible

Structural Balance

- Four types of undirected signed triads
  - Balanced – stable, tend to occur and persist
    - Friend of my friend is my friend (3+)
    - Friend of my enemy is my enemy (1+)
  - Unbalanced – unstable, tend not to occur
    - Friend of my enemy is my friend (2+)
    - Enemy of my enemy is my enemy (0+)

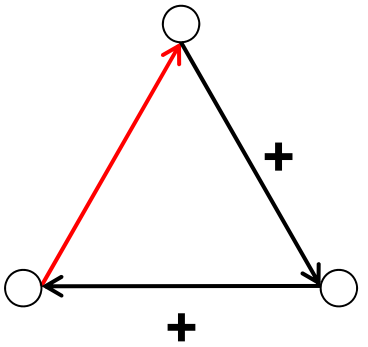


Evaluating Structural Balance

- Expect balanced triads more prevalent and unbalanced less prevalent
  - With respect to baseline based on overall fraction of positive and negative edges
- In all three datasets (epinions, slashdot, wikipedia)
  - 3+ triads massively overrepresented
  - 2+ triads massively underrepresented
  - Mixed results for 0+, 1+ triads
- More consistent with weak balance, only postulates 2+ case unbalanced [Davis 67]
  - Little previous evidence for distinguishing

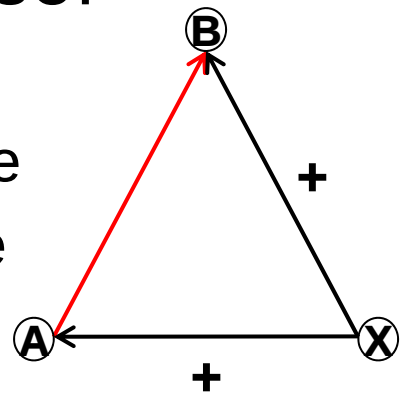
Meanings of a Link

- Edges in these networks actually directed, though direction ignored in balance theory
  - Symmetric relations such as trust, friendship
- Support for weak balance confirms links can reflect such assessments
- But also reflect assessments such as status
  - Asymmetric: generator views recipient as higher (+) or lower (-) status
- Example of cycle with + + path
  - Balance predicts +
  - Status predicts -



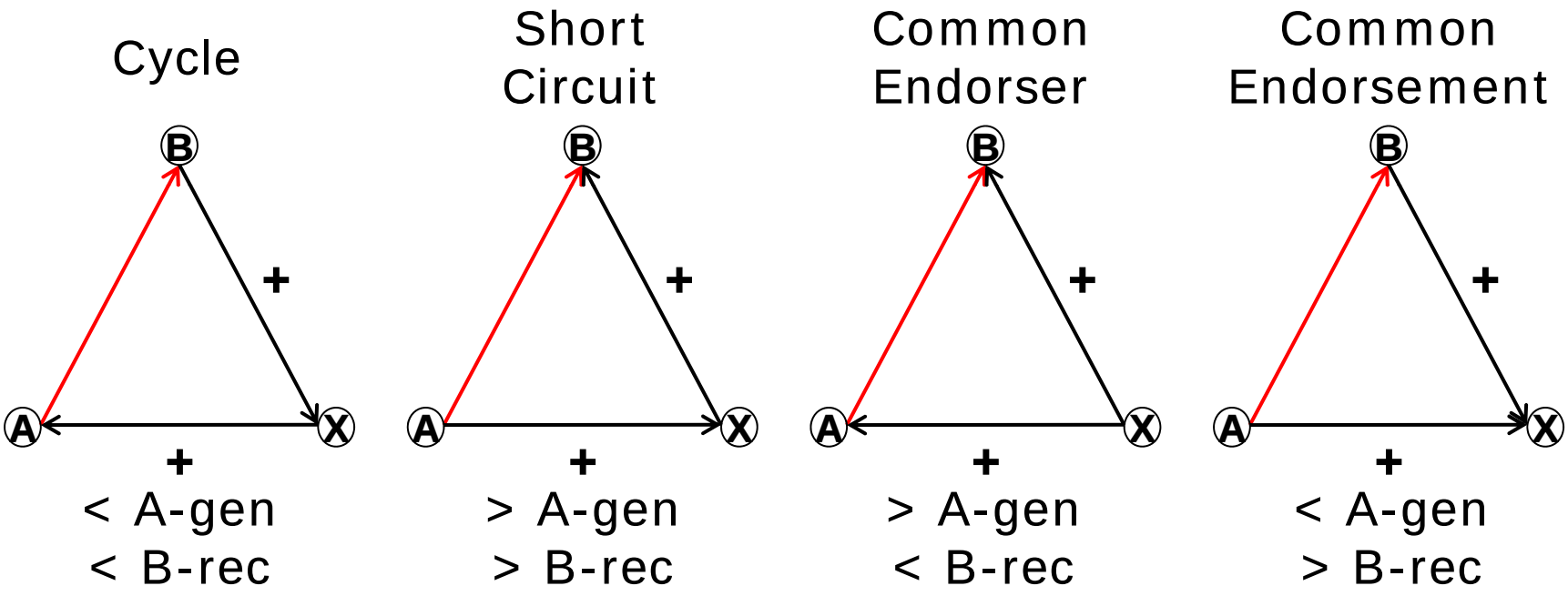
Towards a Theory of Status

- Generative and receptive baselines
  - Higher status individuals will tend to generate negative links and receive positive ones
  - Conversely for a lower status individual
- Edge in given context overrepresented or underrepresented vs. these two baselines
- Example: common positive endorser
  - For pairs of nodes
    - A high status, low generative baseline
    - B high status, high receptive baseline
  - Prediction: > A-gen, < B-rec



Status Theory

- 16 cases – 4 with all positive edges



- Predictions match Wikipedia data in all 4 cases, Epinions in all but A-gen for last case

Comparing Balance and Status

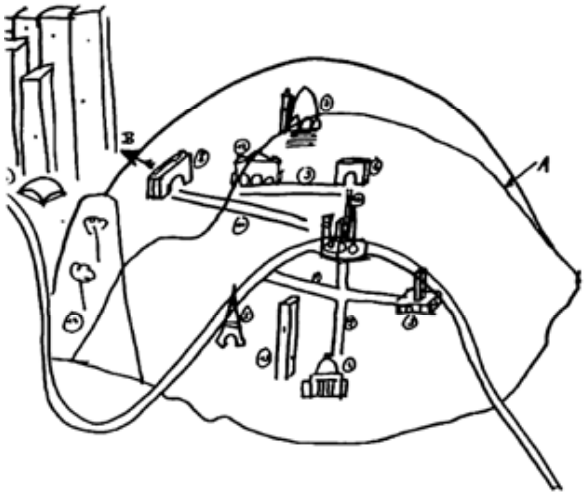
- Status makes correct predictions in
  - 27 of 32 cases for Epinions data
  - 24 of 32 cases for Wikipedia data
- Balance only correct in less than half the cases
- Incorrect predictions for status are mainly when A,B both low-status compared to X
  - Why assessments of and by lower status individuals more difficult?
- Balance better for reciprocated positive edges – but only true for 3-5% of edges

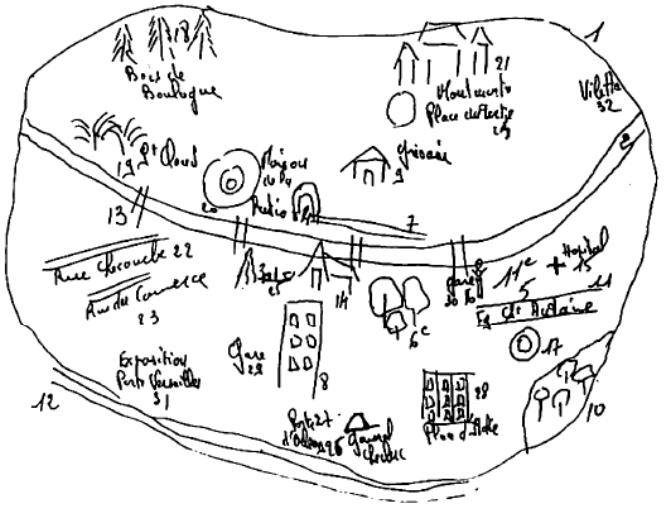
Part II Recap: Signed Edges

- Investigate models that predict prevalence of different forms of triadic relationships
  - Balance and status
- Suggest that signed edges often reflect assessments of status rather than friendship, trust or affinity
- Friendship better explains observed patterns for reciprocated edges
  - But only a small fraction of overall links in these datasets

Part III: Shared Perception of Places

A city is a social fact. We would all agree to that. But we need to add an important corollary: the perception of a city is also a social fact, and as such needs to be studied in its collective as well as its individual aspect. It is not only what *exists* but what is *highlighted* by the community that acquires salience in the mind of the person. A city is as much a collective representation as it is an assemblage of streets, squares, and buildings.

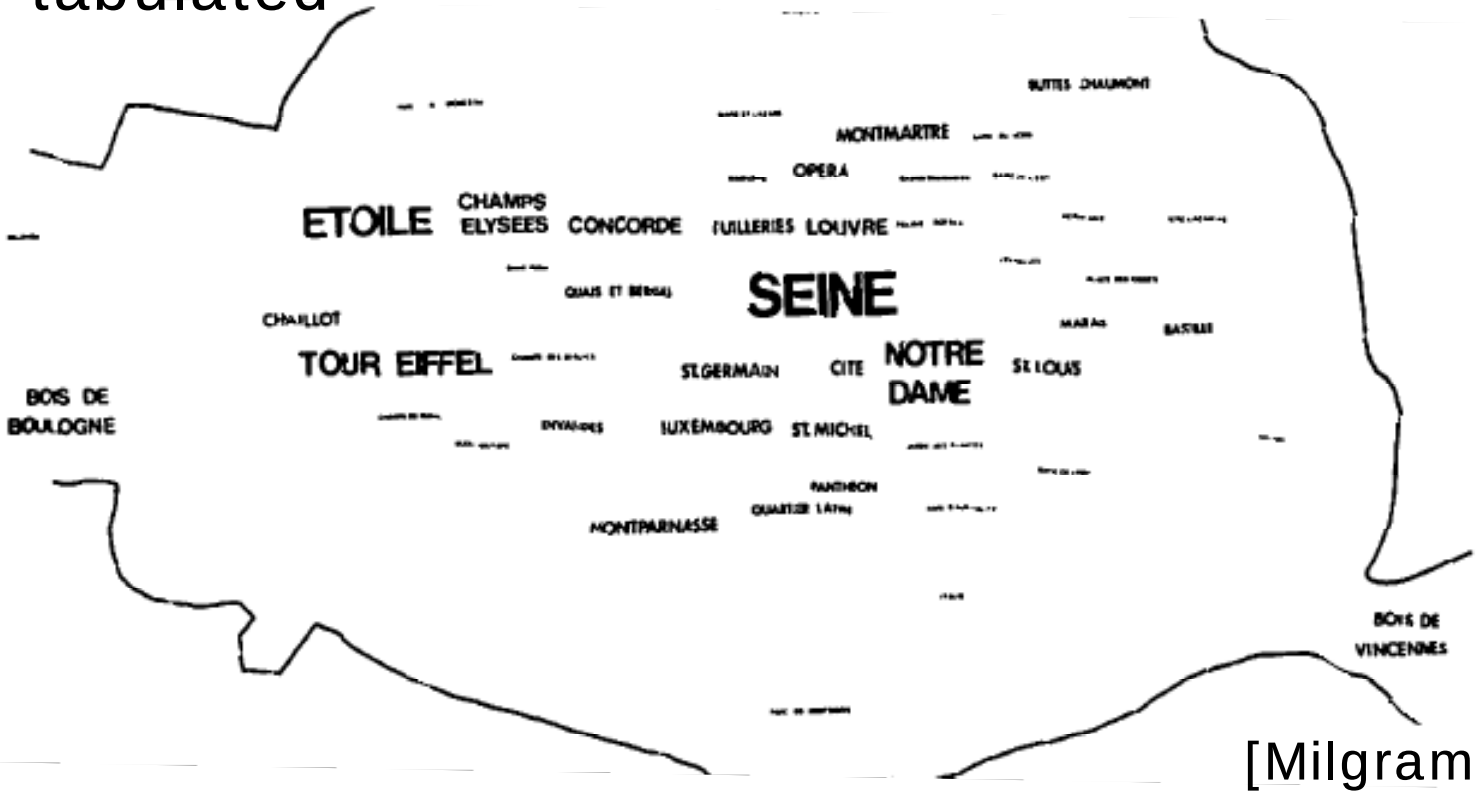




[Milgram76]

Classic Experiment: Hand-Drawn Maps

- 218 subjects each draw map of Paris
  - Total of 4132 elements in maps, hand coded and tabulated



[Milgram76]

Shared Perception in Internet Age

- Billions of publicly available photos online
  - Most with tags – but only somewhat descriptive
  - Hundreds of millions with geo location
    - Growing quickly with new devices
- Large-scale data about the world – extract shared mental maps: places and events
  - From scale of a single city to the globe
  - From hundreds of people to hundreds of thousands or millions
  - From explicit experimental settings to everyday activities

Finding Important Locations

- Natural scales of interest ("octaves")
  - 100km city/metro area, 10km town, 1km neighborhood, 100m landmark
- Want to discover locations automatically at one or more spatial scales
  - Think of geo-tags as samples from unknown spatial distribution whose modes we want to estimate at certain scales
- Mean-shift, procedure for estimating modes
  - Fixed-scale clustering, rather than k-means or agglomerative methods

Sample Clustering Result

- Top 100 clusters in North America at 100km radius – with photos plotted as dots

Representative Text Tags

- Text tags that are characteristic of a given spatial region
  - Score tags according to likelihood in region versus baseline occurrence

$$\frac{P(\text{photo } p \text{ has tag } t \mid p \text{ inside region})}{P(\text{photo } p \text{ has tag } t)}$$

  - Limit any single user's contribution in a region
  - Consider tags that occur for at least some fraction of photos in region (e.g., 5%)
  - Similar approaches in [Ahern07] [Kennedy08]
- Top scoring tags ordered by likelihood

Tags of Top Clusters (100km and 100m)

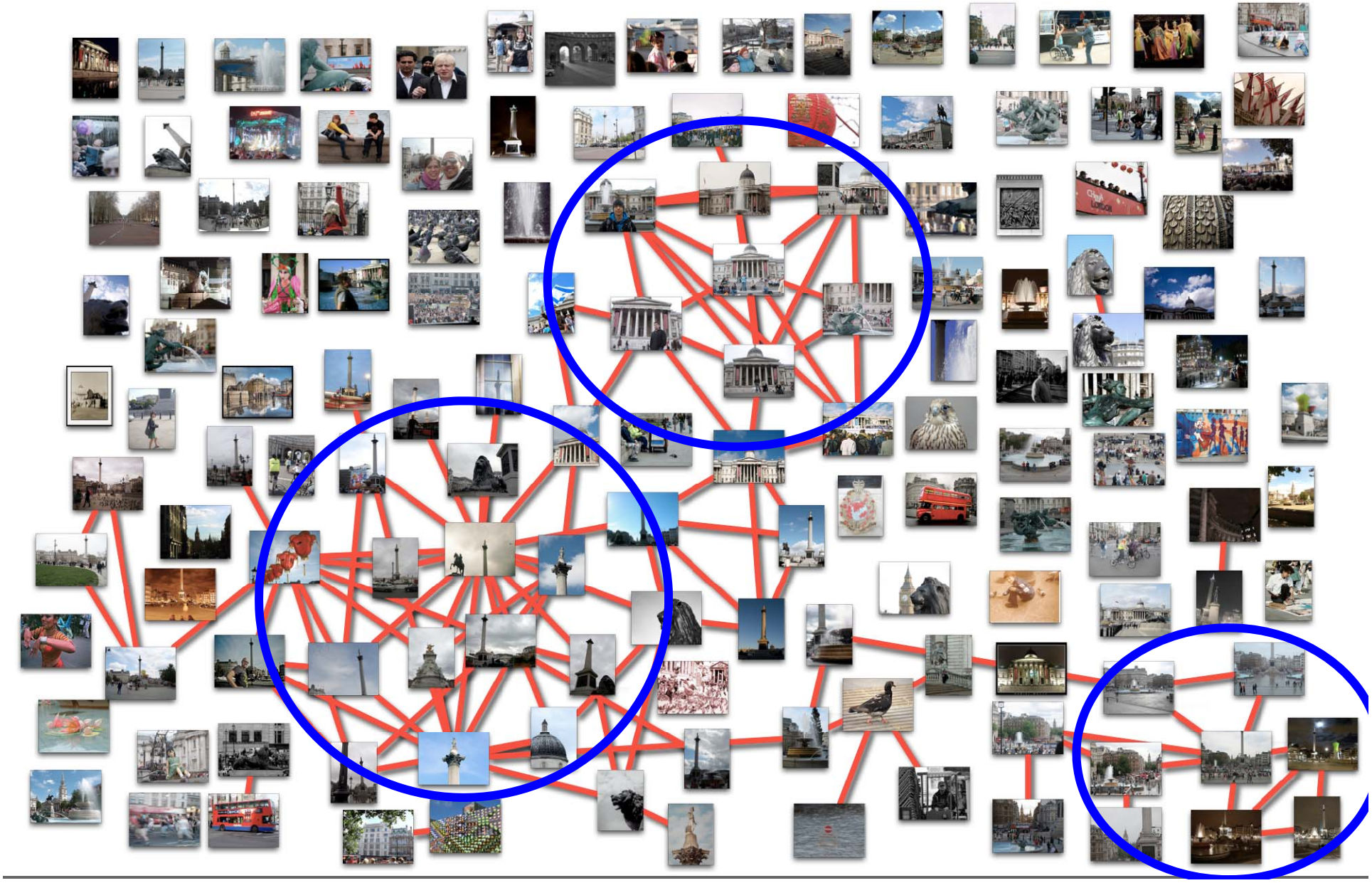
|| Top landmark | 2nd landmark | 3rd landmark | 4th landmark | 5th landmark |

	Top landmark	2nd landmark	3rd landmark
Earth	eiffel	trafalgarsquare	tatemodern
1. **newyorkcity**	empirestatebuilding	timessquare	rockefeller
2. **london**	trafalgarsquare	tatemodern	bigben
3. **sanfrancisco**	coittower	pier39	unionsquare
4. **paris**	eiffel	notredame	louvre
5. **losangeles**	disneyland	hollywood	gettymuseum
6. **chicago**	cloudgate	chicagoriver	hancock
7. **washingtondc**	washingtonmonument	wwii	lincolnmemorial
8. **seattle**	spaceneedle	market	seattlepubliclibrary
9. **rome**	colosseum	vaticano	pantheon
10. **amsterdam**	dam	westerkerk	nieuwmarkt
11. **boston**	fenwaypark	trinitychurch	faneuilhall
12. **barcelona**	sagradafamilia	parcguell	boqueria
13. **sandiego**	balboapark	sandiegozoo	ussmidway
14. **berlin**	brandenburgertor	reichstag	potsdamerplatz
15. **lasvegas**	paris	newyorknewyork	bellagio

Representative Images

- Finding visual characterizations of clusters
  - Harder than selecting high likelihood text tags
  - Similar images primarily when taken at nearly the same place – 100m scale
    - Though some characteristic images at city scale too such as NYC yellow cabs, London buses
  - Similar images are generally a relatively small percentage of all images in a spatial cluster
    - E.g., random photos of (or just near) Independence Hall vs. canonical view such as full facade

Image Similarity Graph in Geo Cluster

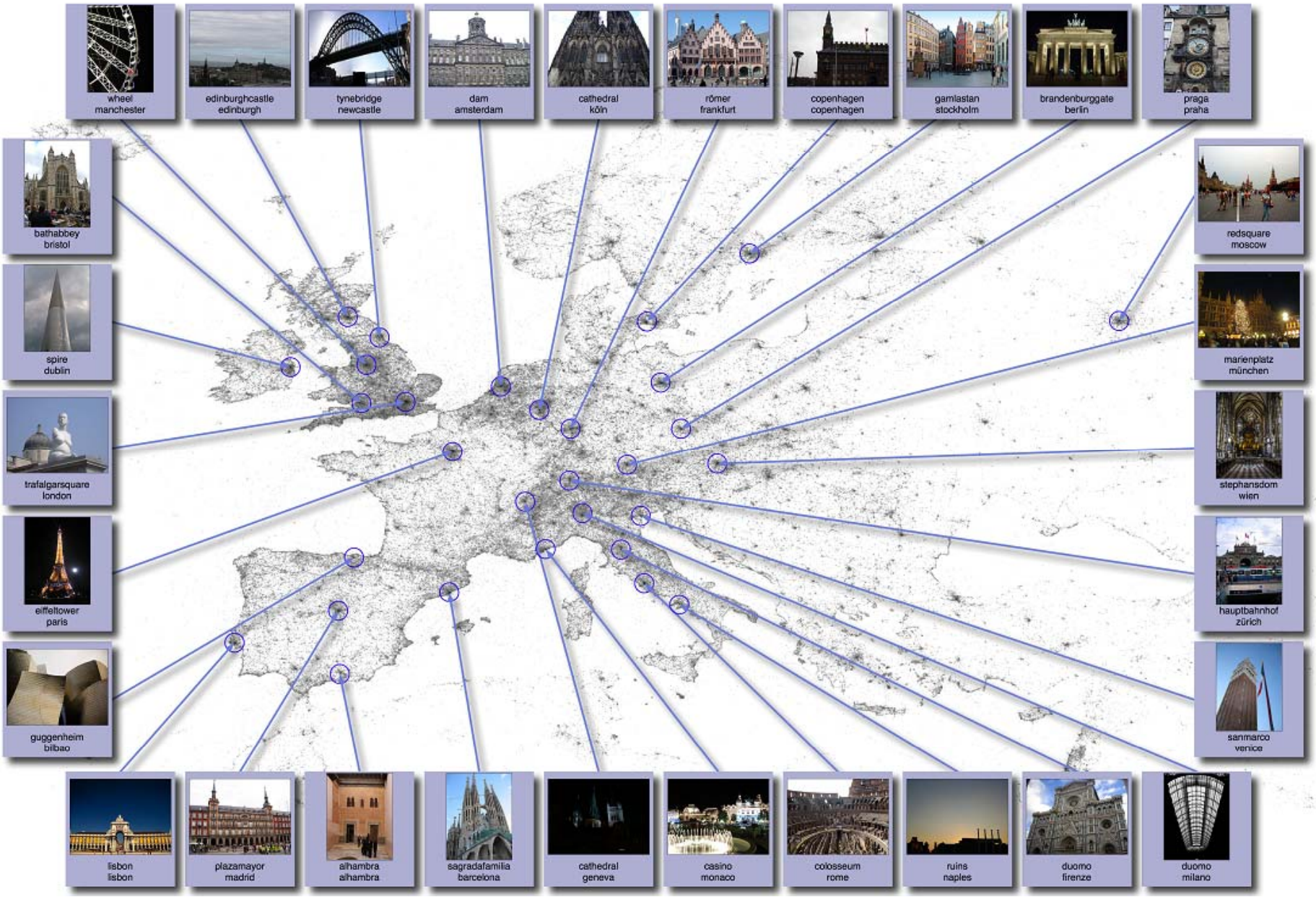


Visualizing Shared Mental Maps

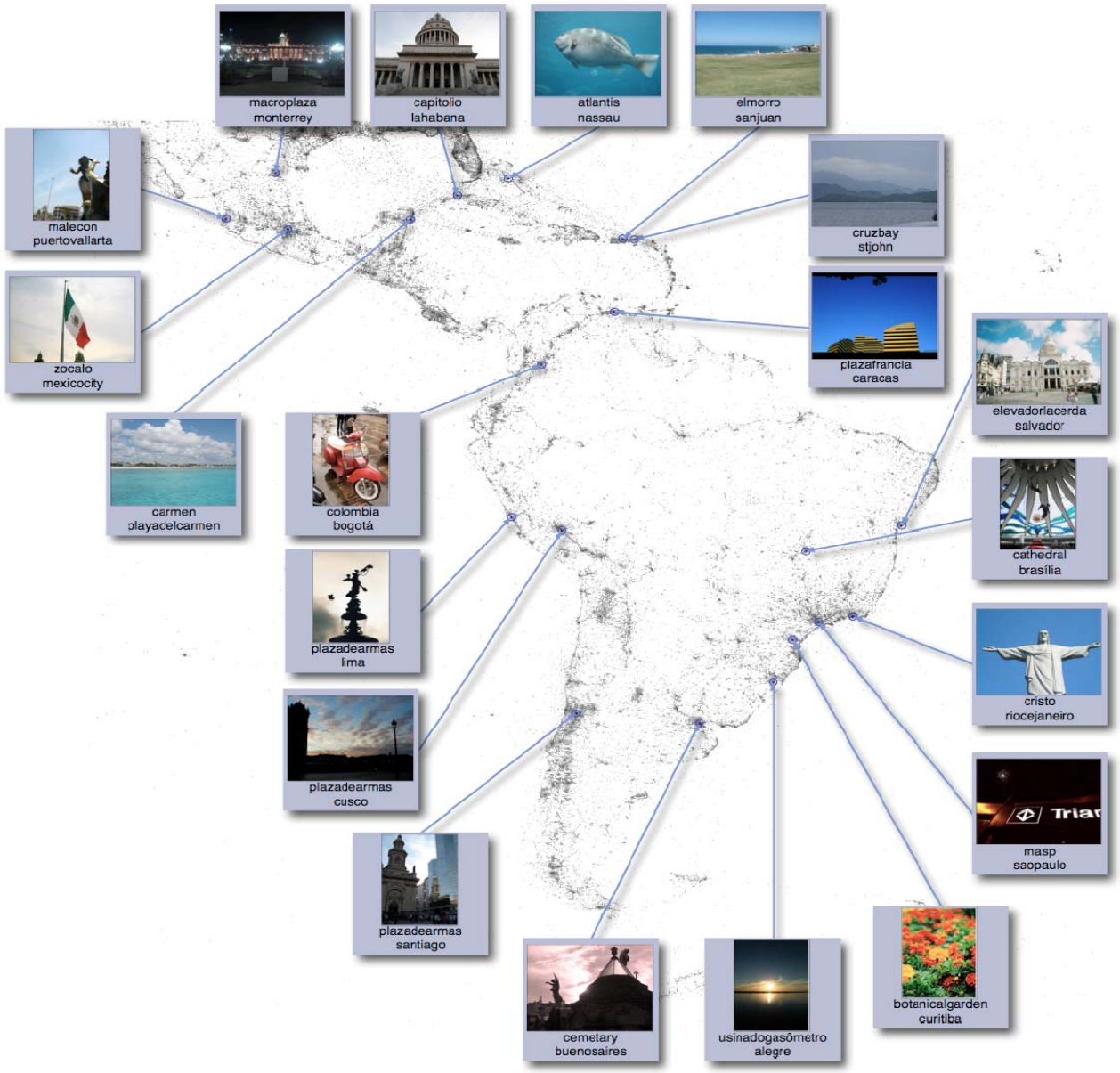
- Scalable techniques for
  - Finding highly-photographed spatial regions, at multiple scales
  - Finding representative textual tags
  - Finding representative images at landmark scale
- Create labeled maps of "what's important" completely automatically
  - City and landmark scales (100km and 100m)
  - From ~35M geo-tagged photos on Flickr, downloaded via API, medium res. (~500 x 350)
- Computation on 400 core Hadoop cluster

Example: Cities in Continental US

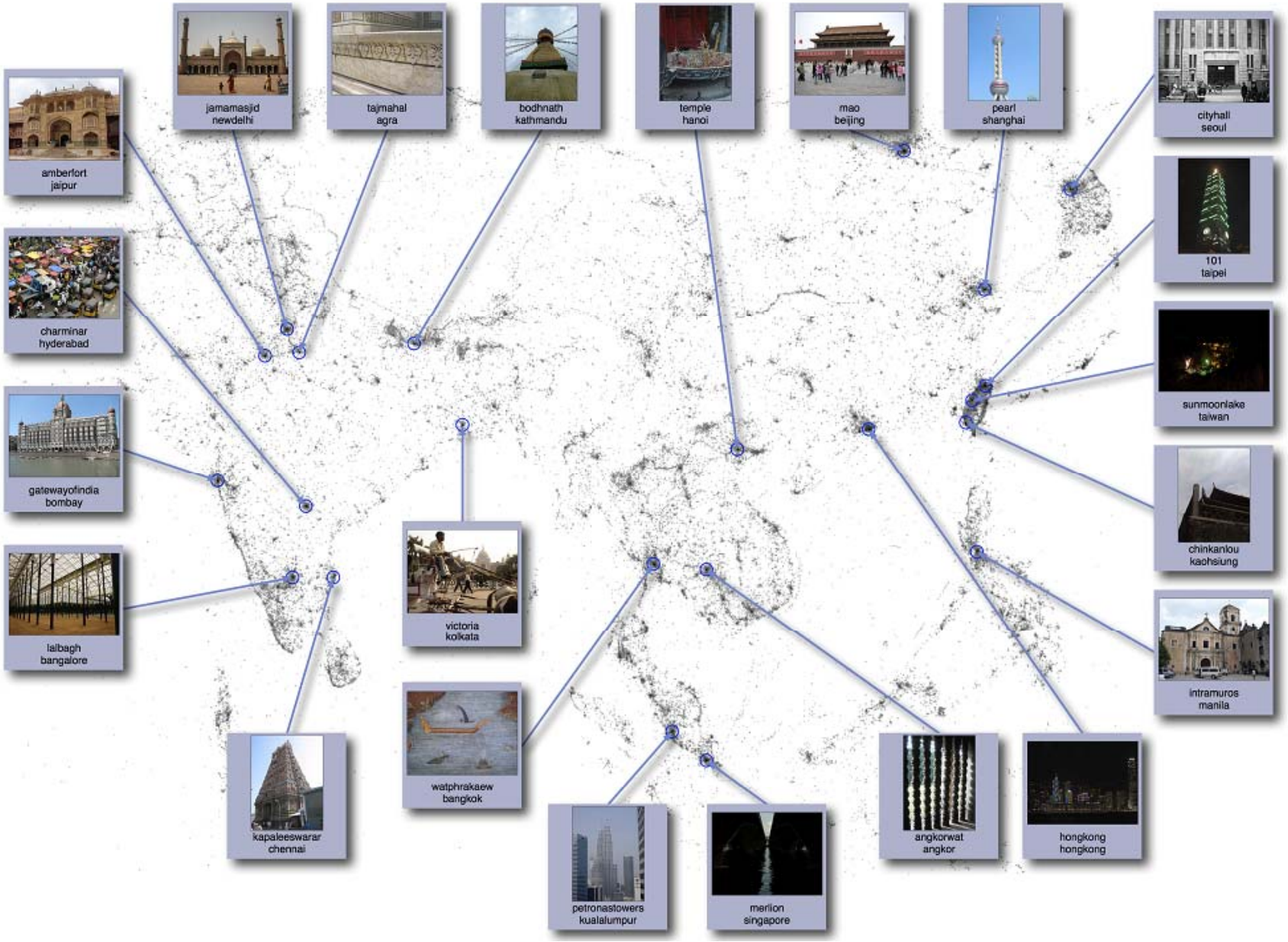
Example: Cities in Europe



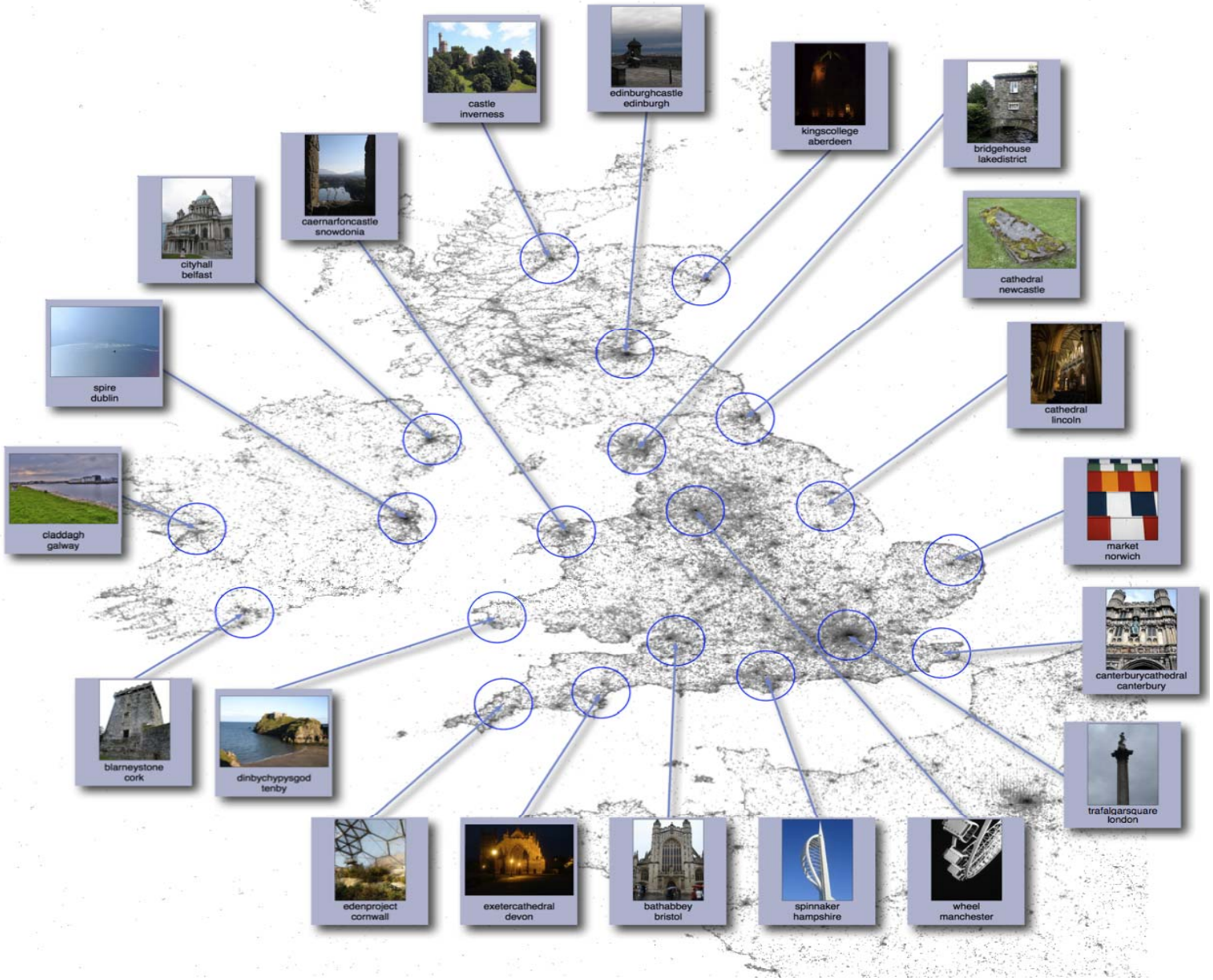
Example: Cities in South America



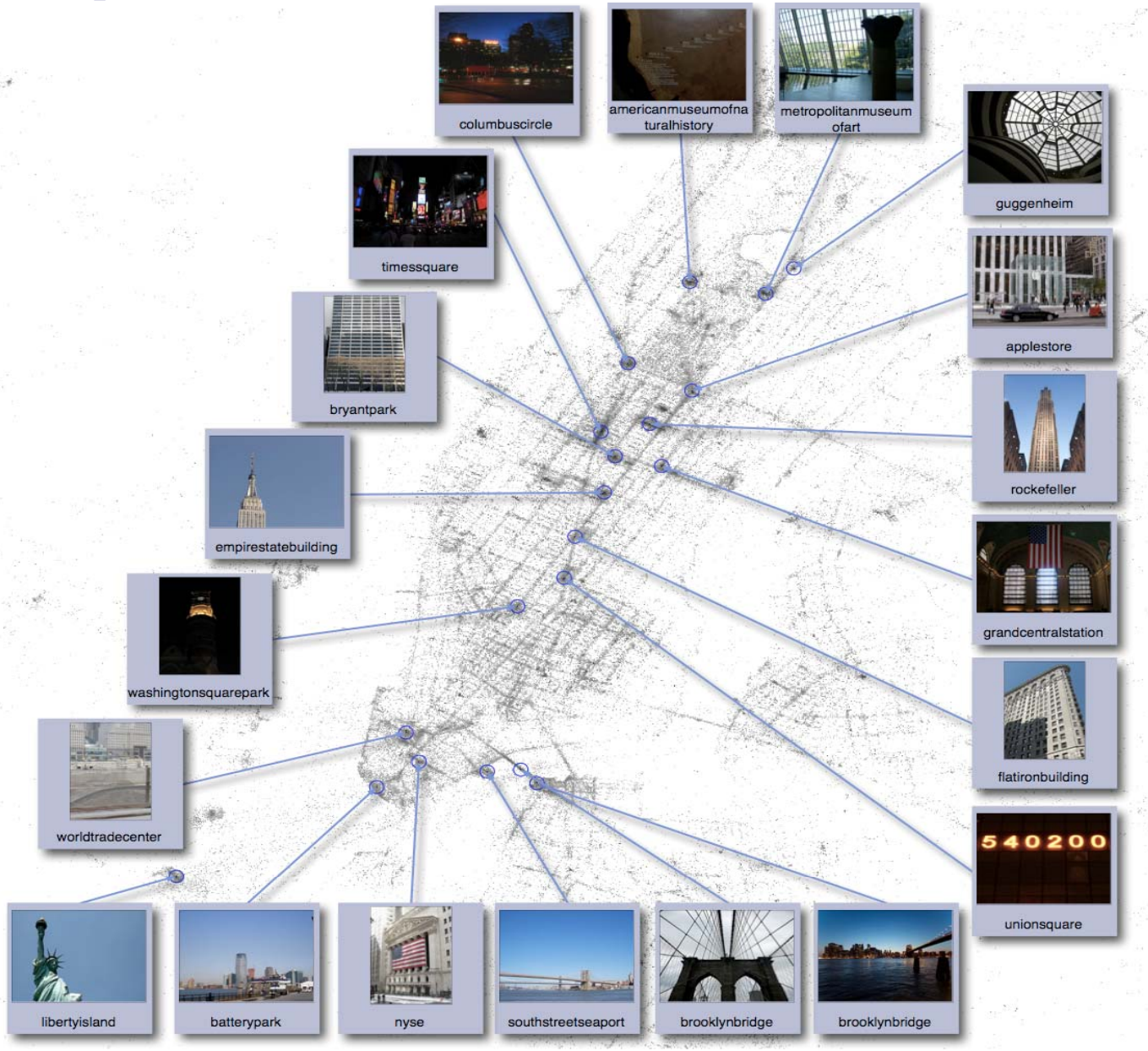
Example: Cities in Southeast Asia



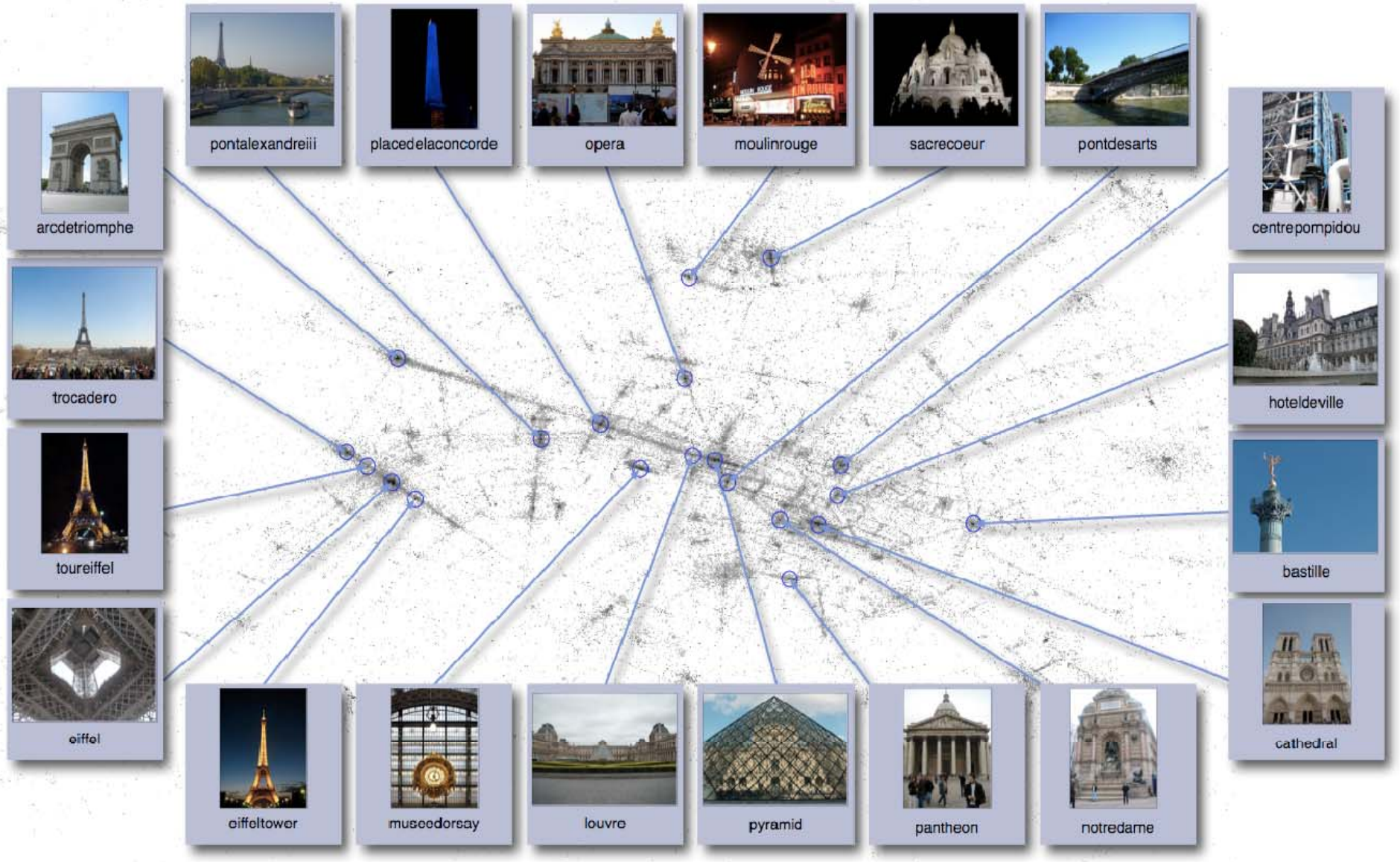
Example: Cities in UK and Ireland



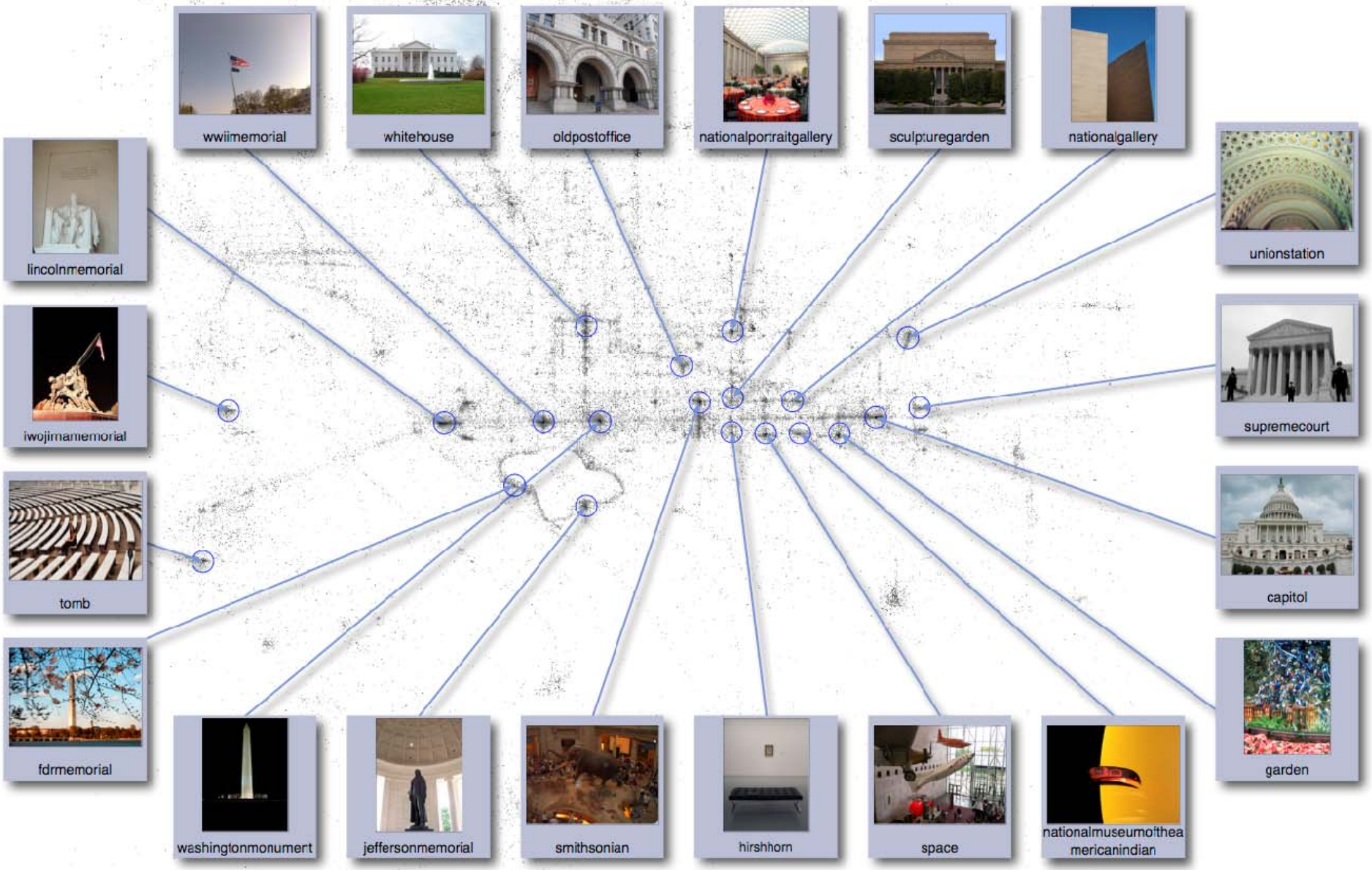
Example: Landmarks in Manhattan



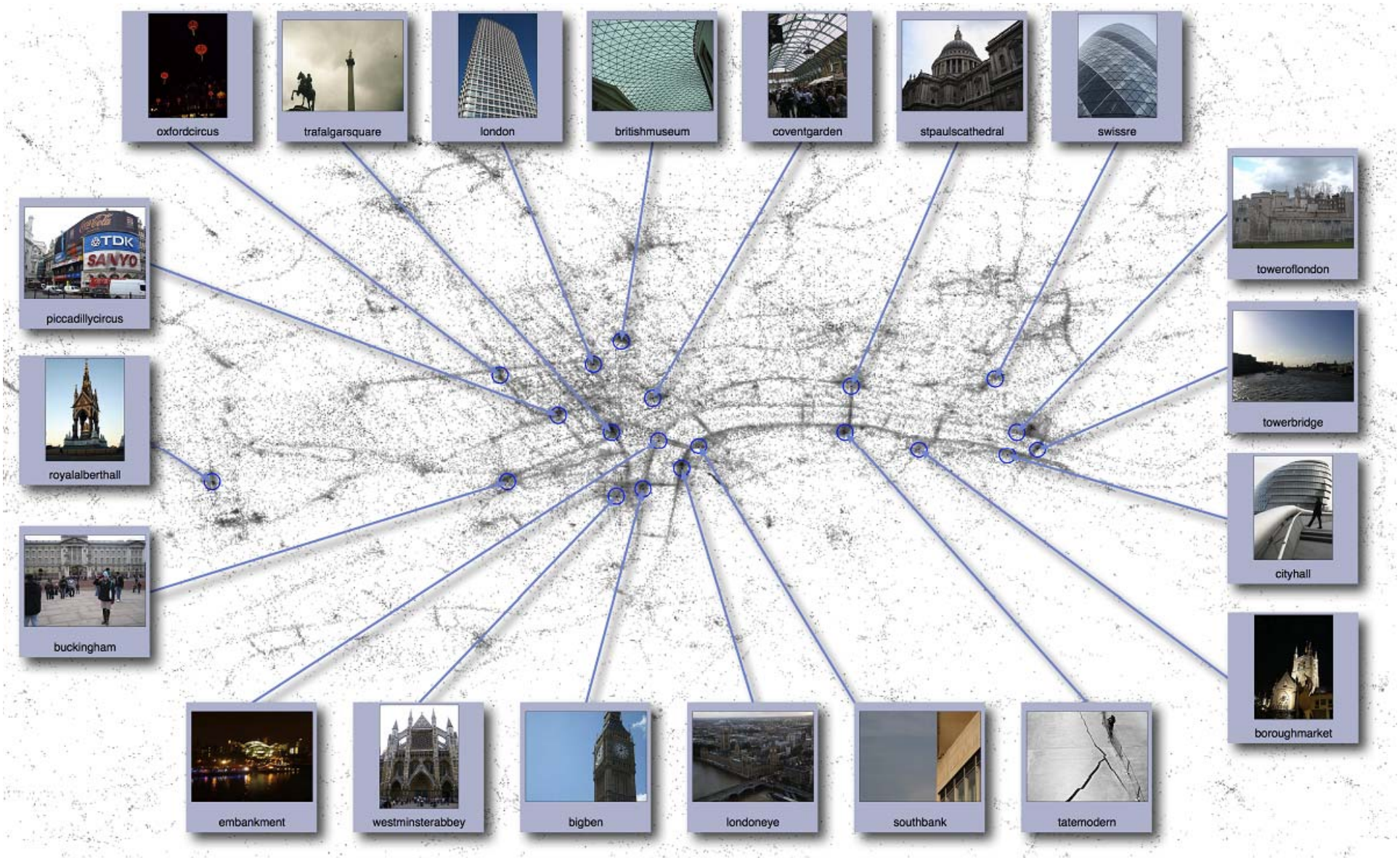
Example: Landmarks in Paris



Example: Landmarks in DC



Example: Landmarks in London



Saliency: Places and Events

- Key aspect of photographic data is that it captures something noteworthy
  - In contrast to always-on location data such as cell phone tracks
- But this high value data can also be very revealing of other (unintended) information
  - People who take photos nearby in time and location generally know each other
    - E.g., if two Flickr users take photos in 6 places within 24 hours and 0.1 degrees (~8km) there is nearly 70% probability they are listed as contacts!
      - Chance only .005% (moreover contacts underreport)

Part III Recap: Shared Perception

- Places that people choose to record photographically serve as a form of link
  - Shared record of what is important or interesting – automated maps of world
- Reflect aspects of human activity and interests that previously were hard to investigate experimentally
  - E.g., city planning interest in socially generated photo maps
- Highly correlated with explicit forms of social linking

In Conclusion

- Socially generated data is opening up a host of research problems, both old and new
  - Calls for methodologies bridging social sciences with computing and information sciences
- Fundamental questions about social interactions and behavior
  - Online and possibly offline
- Practical implications for design and development of social media sites
- Goal of generally applicable scientific tools for social data modeling and analysis

Collaborators

- This talk covers material from papers in KDD06, KDD08, WWW09, ICCV09 and forthcoming in CHI. It is joint work with
  - Lars Backstrom (Facebook)
  - Dan Cosley (Cornell)
  - David Crandall (Cornell)
  - Jon Kleinberg (Cornell)
  - Xiangyang Lan (Goldman)
  - Jure Leskovec (Stanford)
  - Yunpeng Li (Cornell/EPFL)
  - Sid Suri (Yahoo)

Questions