Group Formation in Large Social Networks: Membership, Growth, and Evolution

Lars Backstrom, Dan Huttenlocher, Joh Kleinberg, Xiangyang Lan

Presented by Dung Nguyen

Based on slide of Natalia :
http://www.cs.kent.edu/~jin/dataminingcourse/PPT/Natalia.ppt

Outline

- Introduction
- Membership, Growth, Evolution
- Conclusions

Introduction

Understand:

- Factors that make a person join in a group
- Which Structure properties influence the growth of a community
- What are under the movement from a community to another community? What’s the effect this movement.

Membership, Growth, Change

- Membership
  - Structural features that influence whether a given *individual will join a particular group*
- Growth
  - Structural features that influence whether a given *group will grow significantly over time*
- Change
  - How *focus of interest changes* over time
  - How these changes are *correlated with changes in the set of group members*

Sources of data

- LiveJournal
  - Free on-line community with ~ 10 mln members
  - 300,000 update the content in 24-hour period
  - Maintaining journals, individual and group blogs
  - Declaring who are their friends and to which communities they belong
- DBLP
  - On-line database of computer science publications (about 400,000 papers)
  - Friendship network – co-authors in the paper
  - Conference - community

Method Description

- Use decision trees to figure out what is the most affected factor.

Community Membership

- Study of processes by which individuals join communities in a social network
- Fundamental question about the evolution of communities: who will join in the future?
- Membership in a community – “behavior” that spreads through the network
  - *Diffusion of innovation* study perspective for this question

Considered factors:

Features related to the community, C. (Edges between only members of the community are $E_C \subseteq E$.)	Number of members ($\|C\|$). Number of individuals with a friend in C (the *fringe* of C) . Number of edges with one end in the community and the other in the fringe. Number of edges with both ends in the community, $\|E_C\|$. The number of open triads: $\|\{(u,v,w)\|(u,v) \in E_C \wedge (v,w) \in E_C \wedge (u,w) \notin E_C \wedge u \neq w\}\|$. The number of closed triads: $\|\{(u,v,w)\|(u,v) \in E_C \wedge (v,w) \in E_C \wedge (u,w) \in E_C\}\|$. The ratio of closed to open triads. The fraction of individuals in the fringe with at least k friends in the community for $2 \leq k \leq 19$. The number of posts and responses made by members of the community. The number of members of the community with at least one post or response. The number of responses per post.

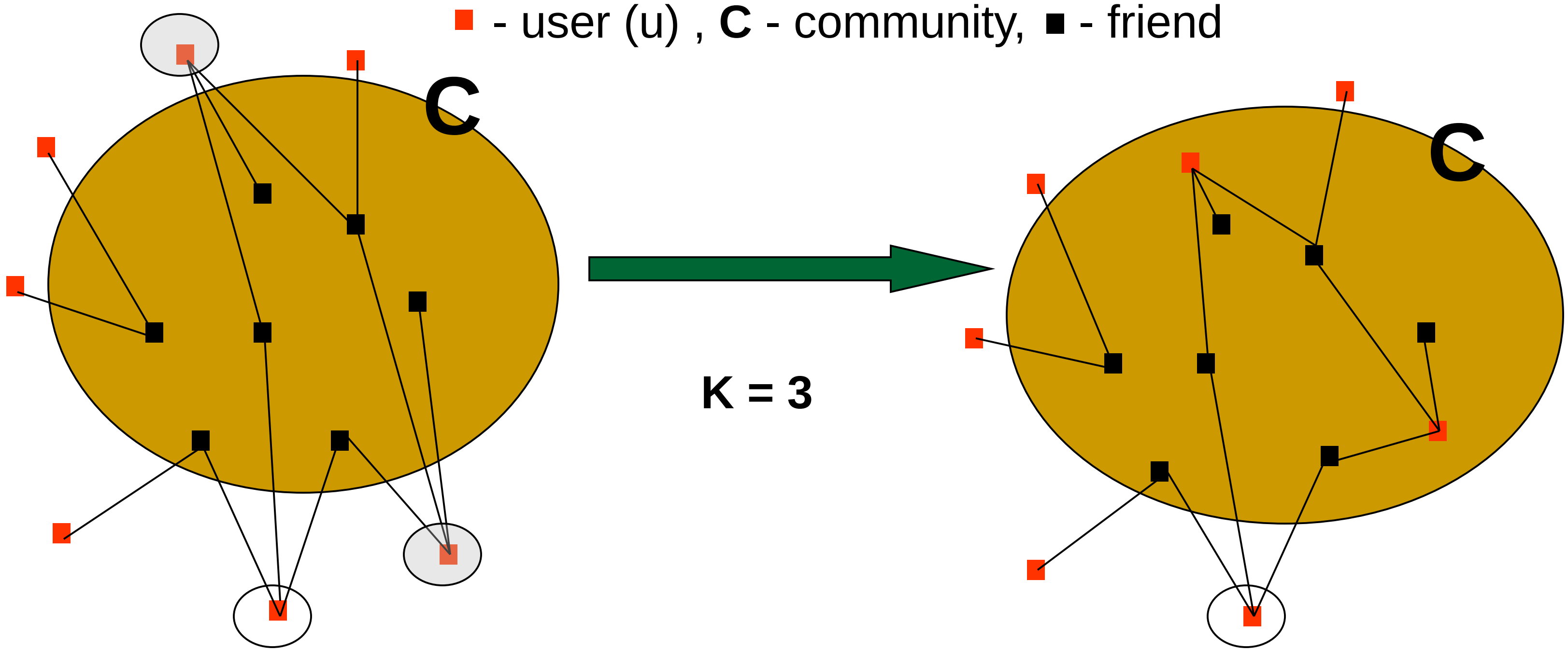
Dependence on number of friends: start towards membership prediction

- Underlying premise in diffusion studies: *an individual probability of adopting a new behavior increases with the number of friends (K) already engaging in the behavior*

- Theoretical models concentrate on the effect of *K*, while the structural properties are more influential in determining membership

Dependence on number of friends

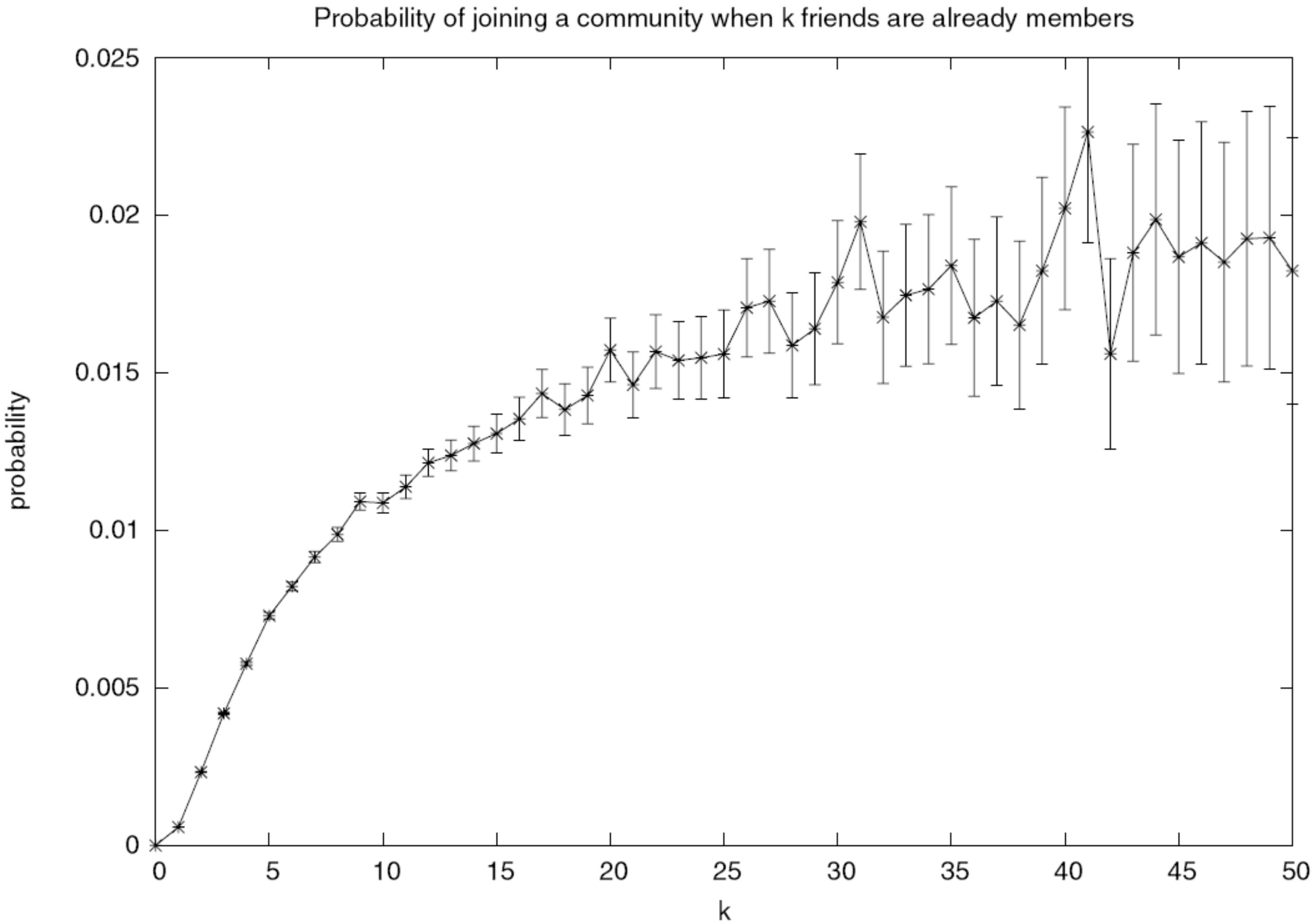
1_{st} snapshot $\qquad$ 2_{nd} snapshot



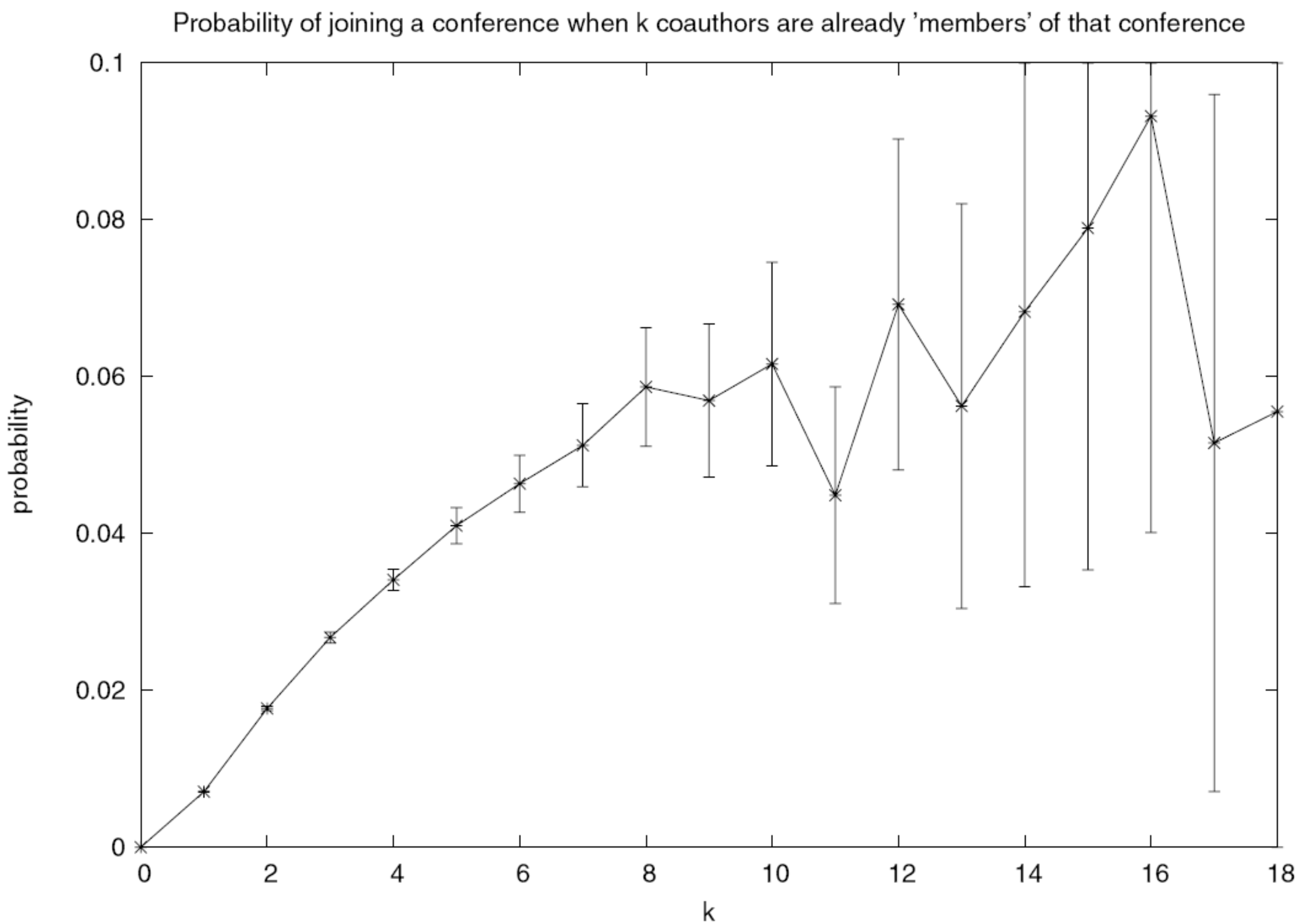
P(k) = 2/3

Probability P(k) of joining community = fraction of triples (u,C,k)

Dependence on number of friends: LiveJournal



Dependence of number of friends: DBLP



More factors:

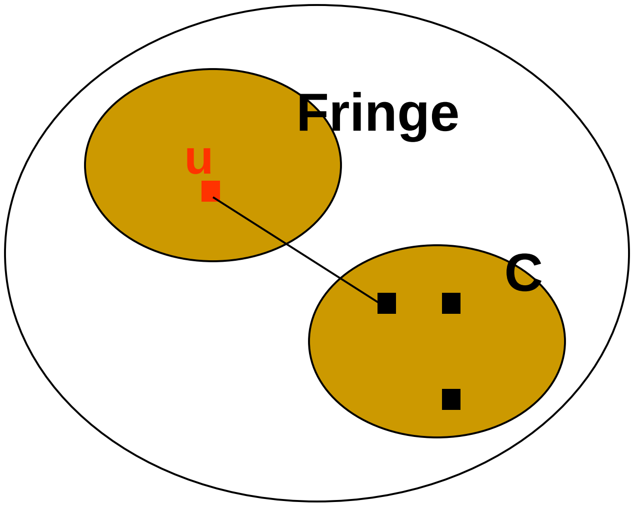
- Features related to the community *C* (11)
  - Number of members *(|C|)*
  - Number of individuals with a friend in *C* ***(fringe of C)***
  - Number of edges with both ends in the community *(|E_C|)*
  - etc.

- Features related to an individual *u* and her set *S* of friends in community *C* (8)
  - Number of friend in community *(|S|)*
  - Number of adjacent pairs in *S*
  - Number of pairs in S connected via a path in E_C
  - etc.

Predictions for LJ and DBLP

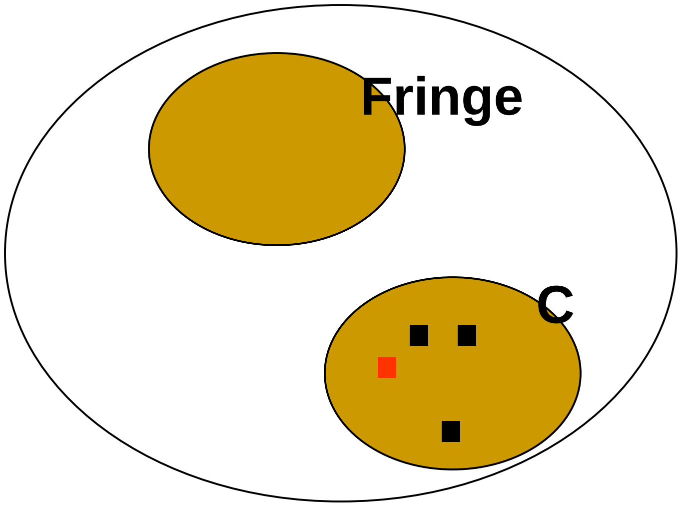
- 1st snapshot

Data point (u,C)



2nd snapshot

Probability $U \in C$

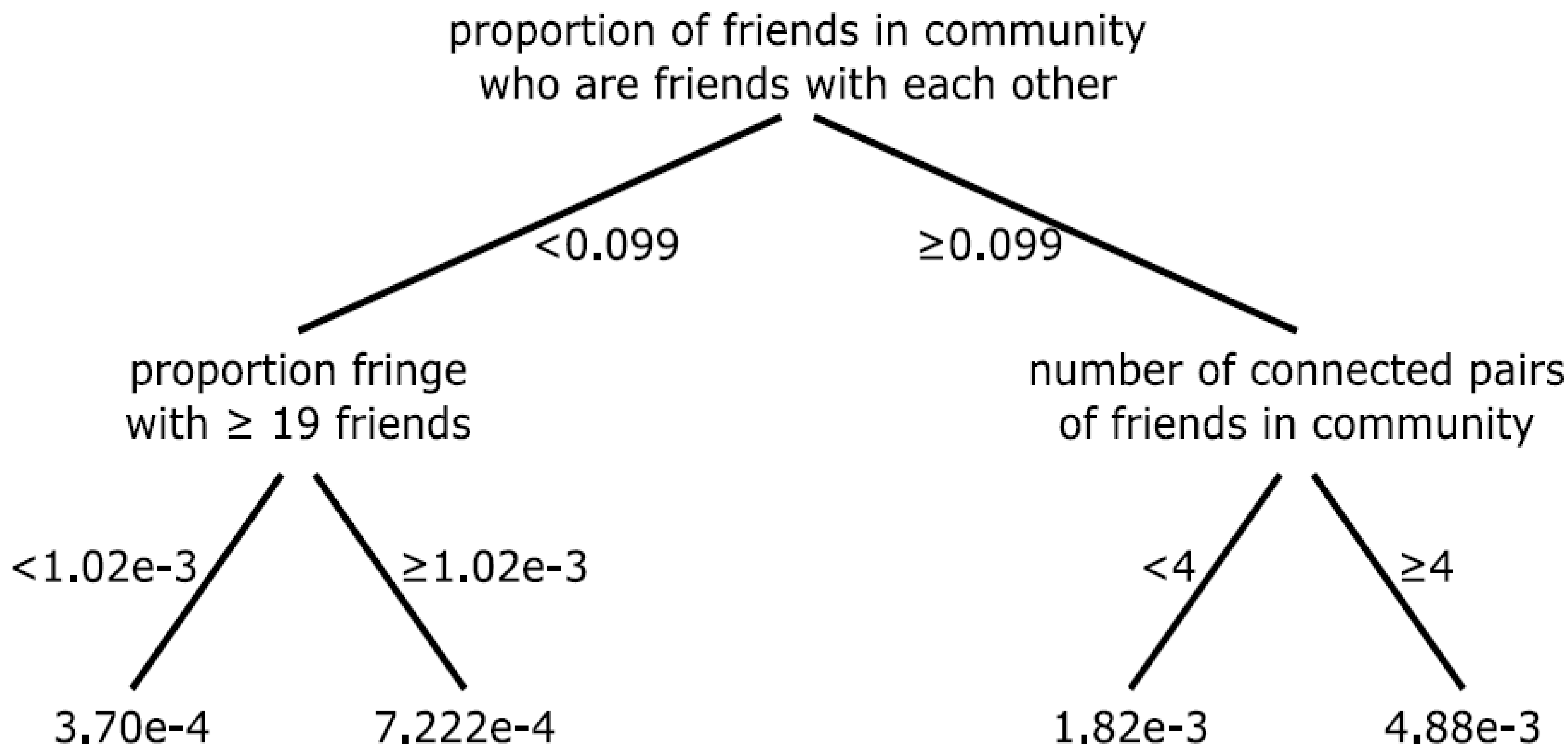


LJ: **17,076,344** data points, **875** communities
DBLP: **7,651,013** data points

LJ: **14,448** joined community
DBLP: **71,618** joined community

20 decisions tree were built for estimation about joining

Top two level splits for predicting single individuals joining communities in LJ



Performance achieved with the decision trees

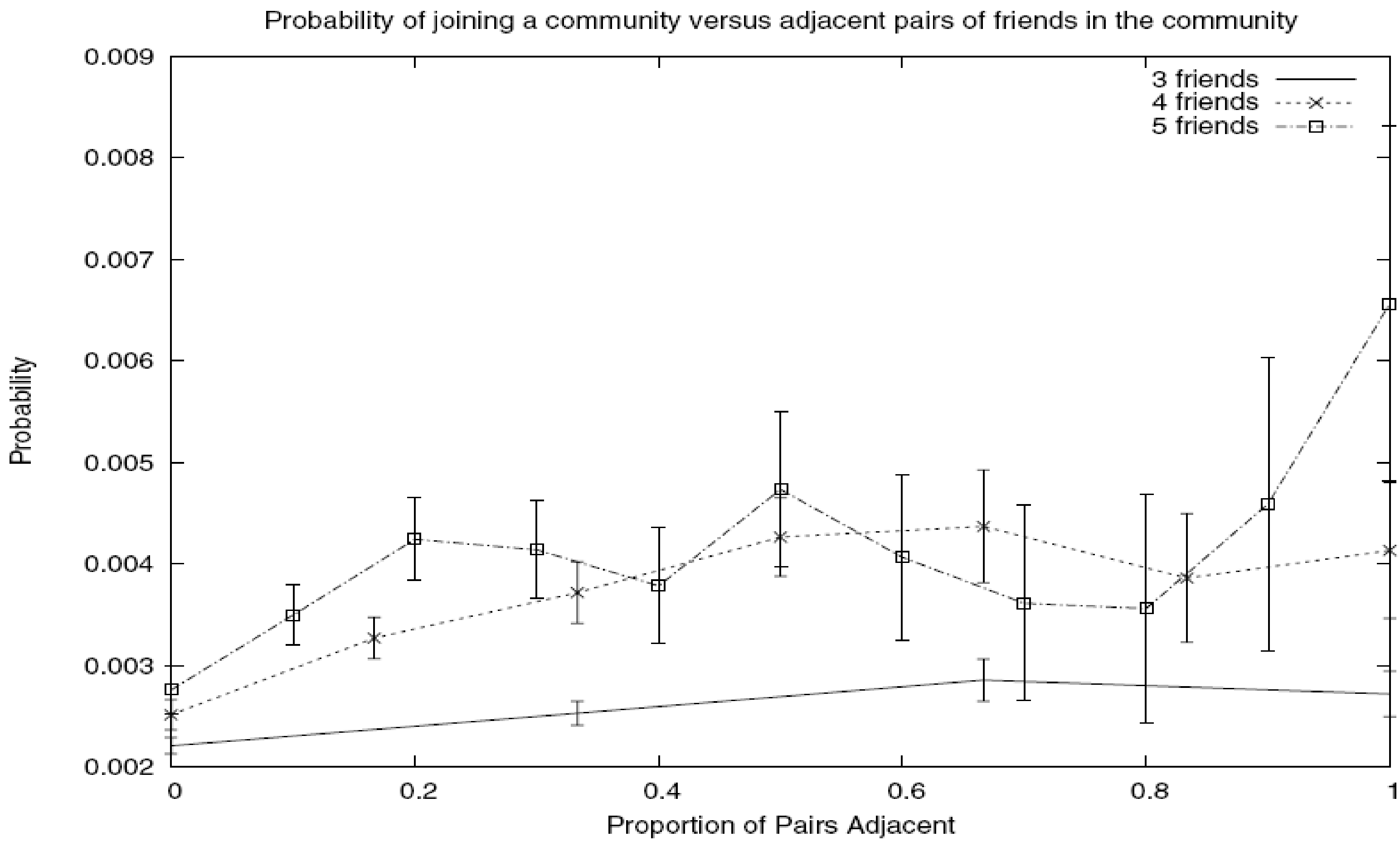
Features Used	ROCA	APR	CXE
Number of Friends	0.69244	0.00301	0.00934
Post Activity	0.73421	0.00316	0.00934
All	0.75642	0.00380	0.00923

Prediction performance for single individuals joining communities in LJ

Features Used	ROCA	APR	CXE
Number of Friends	0.64560	0.01236	0.06123
All	0.74114	0.02562	0.05808

Prediction performance for single individuals joining communities in DBLP

Internal connectedness of friends



Individuals whose friends in community are linked to one another are significantly more likely to join the community

Community Growth

- Three baselines with a single feature were considered
  - Size of the community
  - Number of people in the fringe of the community
  - Ratio of these two features and combination of all three features

Results

Features Used	ROCA	APR	CXE	ACC
Fringe	0.55874	0.53560	1.01565	0.54451
Community Size	0.52096	0.52009	1.01220	0.51179
Ratio of Fringe to Size	0.56192	0.56619	1.01113	0.54702
Combination of above 3	0.60133	0.60463	0.98303	0.57178
All Features	0.77070	0.77442	0.82008	0.70035

Predicting community growth: baselines based on three different features, and performance using all features

By including the full set of features predictions with reasonably good performance were received

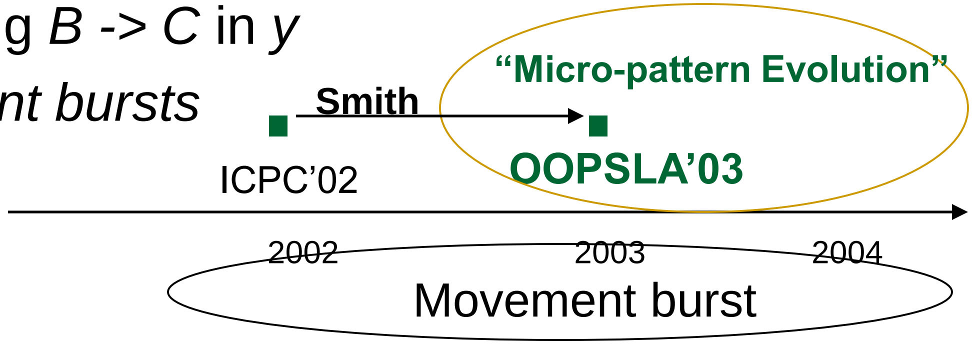
Movement between communities

- How people and topics move between communities
- Fundamental question: given a set of overlapping communities
  - do topics tend to follow people
  - or do people tend to follow topics
- Experiment set up: 87 conferences for which there is DBLP data over at least 15-year period
  - Cumulative set of words in titles is a proxy for top-level topics

Experiment 1: Papers contributing to Movement Bursts

- Characteristics of *papers associated with some movement burst* into a conference *C*
  - They exhibit different properties from arbitrary papers at *C*
    - Using of terms currently hot at *C*
    - Using of terms that will be hot at *C* in the future

- Paper at *C* in *y* *contributes* to some movement burst at C
  - If one of the authors is moving *B* -> *C* in *y*
  - *y* is a part of *B* -> *C movement bursts*



Papers contributing to Movement Bursts

- Paper *uses hot term*
  - If one of the words in its title is hot for the conference and year in which it appears
- Question: do papers contributing to movement bursts differ from arbitrary papers in the way they use hot terms?

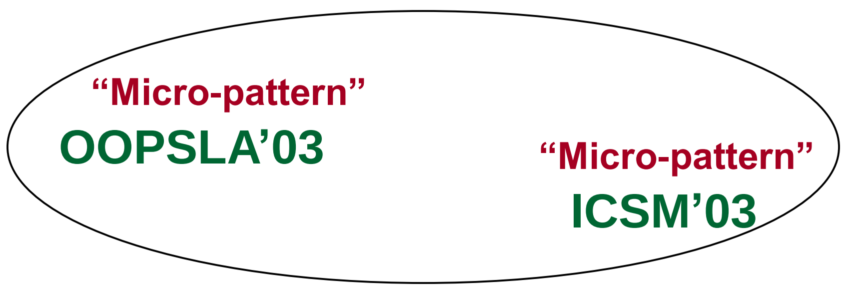
	All Papers	Papers Contrib. to Movement
Num. papers	99774	10799
Currently hot	0.3859	0.4391
Future hot	0.1740	0.1153
Expired hot	0.2637	0.3102

Papers contributing to a movement burst contain elevated frequencies of currently and expired hot terms, but lower frequencies of future hot terms

A burst of authors moving into *C* from *B* are drawn to topics currently hot at *C*

Experiment 2: Alignment between different conferences

- Conferences *B* and *C* are *topically aligned* in a year *y*
  - If some word is hot at both *B* and *C* in year *y*
  - Property of two conference and a specific year



- Hypothesis: two conferences are more likely to be topically aligned in a given year if there is also a movement burst going between them

Results

- 56.34% of all triples *(B,C,y)* such that *there is B->C movement burst containing year y* have the property that *B* and *C* are topically aligned in year *y*

- *16.2 %* of *all* triples *(B,C,y)* have the property that *B* and *C* are topically aligned in year *y*

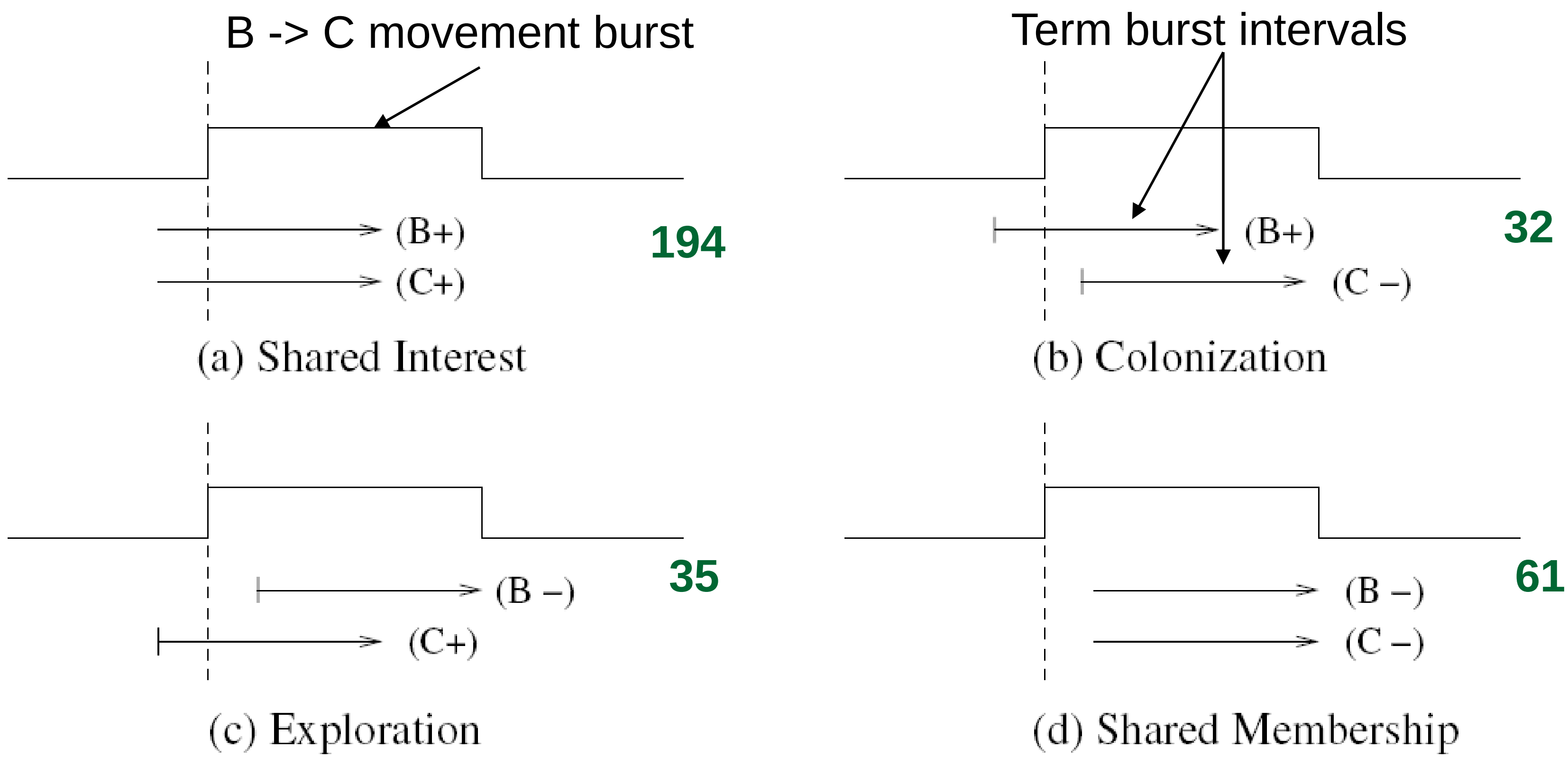
- The *presence of a movement burst* between 2 conferences enormously increases the chance they share a hot term

Movement bursts or term bursts come first?

- There is a *B -> C movement burst,* and hot terms w such that *B* and *C* are topically aligned via *w* in some year *y* inside the movement burst

- 3 events of interest
  - The start of the burst for *w* at conference *B*
  - The start of the burst for *w* at conference *C*
  - The start of the B -> C movement burst

Four patterns of author movement and topical alignment



Shared interest is 50 % more frequent than others

Much more frequent for *B* and *C* to have a shared burst term that is already underway before the increase in author movement takes place

Conclusions

- Heuristic predict the change of community.
- Remodel the problem “information diffusion”
- Problem: how to grow a community with limited budget?
- Problem: how to attack other community with limited budget?

Thank you!