Applying Link-based Classification to Label Blogs

Graham Cormode

Smriti Bhagat, Irina Rozenbaum

Blogs as Multigraphs

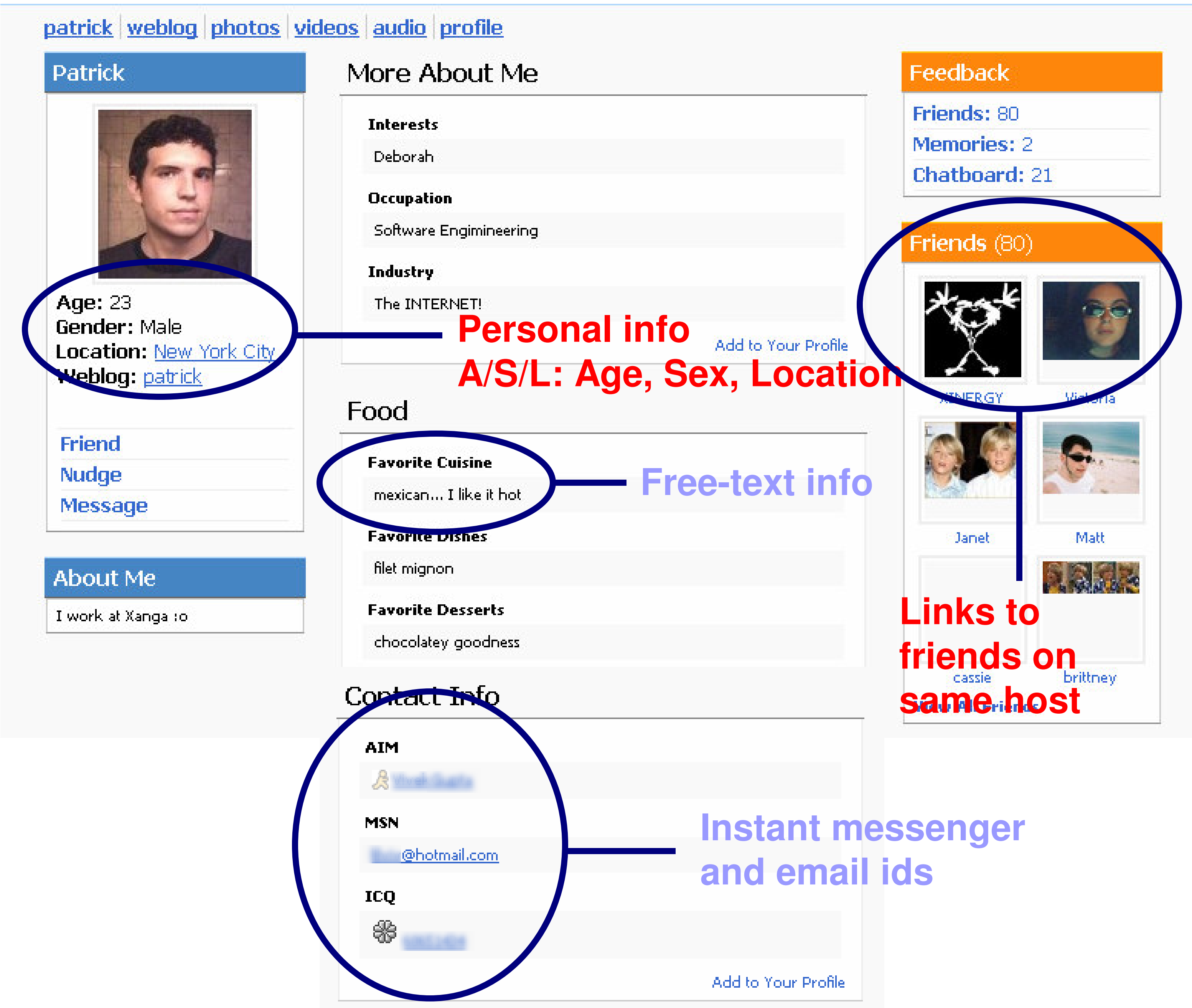
Many interesting new data sources are best modelled as multigraphs, with multiple attributes and link types.

“Blogs” are an important emerging example of such data:

- Intersect with web, email, chat data, social networks
- React rapidly to major news, defining opinion and identifying articles of interest
- Raise problems of trustworthiness, finding leaders, classifying for expertise and bias

We study labeling problems on these large multigraphs

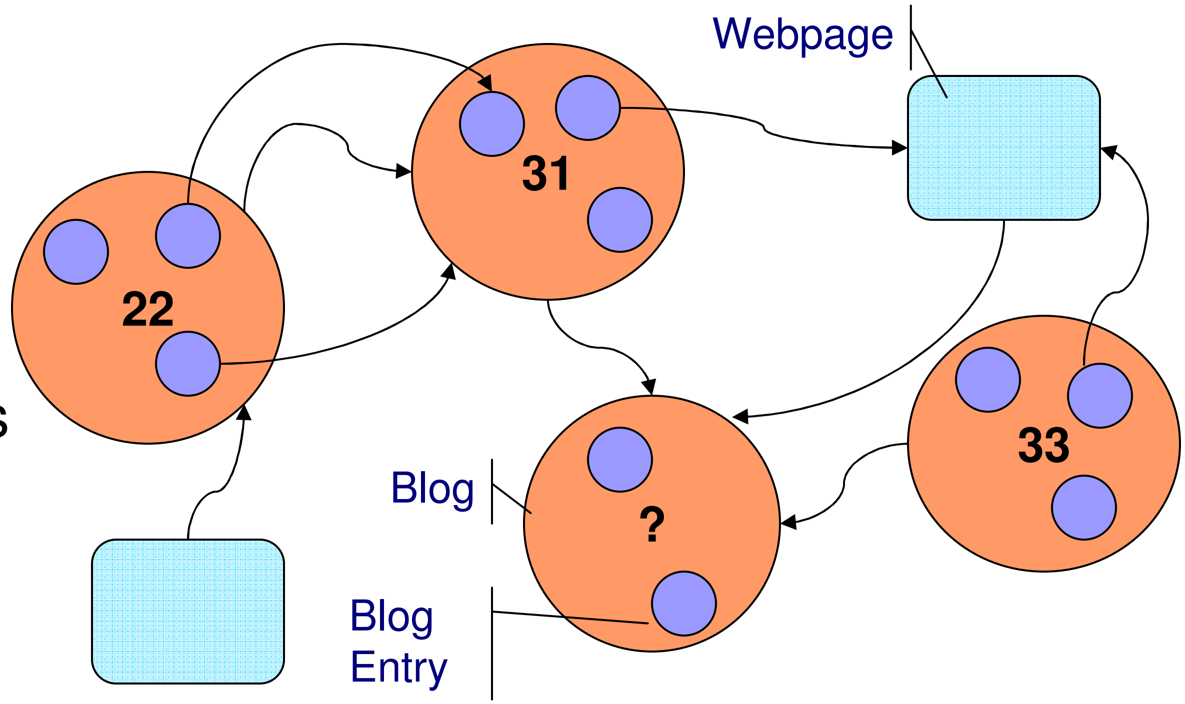

THE GEOMBLOG
RUMINATIONS ON COMPUTATIONAL GEOMETRY, ALGORITHMS, THEORETICAL COMPUTER SCIENCE AND LIFE
ABOUT ME
VIEW MY COMPLETE PROFILE
profile data
» Blogs that link here
Technorati
SITE METER
251,100
QUICKLINKS
» The Matrix Cookbook
» STXXL: C++ Library for Extra Large Data Sets
» Doron Zeilberger RSS feed
» John Baez RSS feed
» Halting Problem Poem: Scooping The Loop Snooper
TUESDAY, AUGUST 15, 2006
timestamp
Poincare's conjecture
headline
does an excellent job describing both the conjecture itself and the main line of attack that Perelman
follows, while narrating the history of the proof both accurately and without unnecessary romanticization
of the mathematicians involved.
text
For a more detailed nontechnical survey that goes more into the roles of the different players, see Allyn
Jackson's article on the latest Notices of the AMS. A more technical account is
manuscript by Shing-Tung Yau.
links
Categories: research
author tags
2 COMMENTS:
reader comments
http://www.newyorker.com/fact/content/articles/060828fa_fact2
Posted by Dmitry Sustretov
Commenter id and timestamp
8/27/2006 02:05:26 PM
Poincare conjecture is very important for mathemetical and phyisical world.The proof of this
conjecture,doubtless, will bring to science world new thinks and approachs. we saw in history the
similarities of this. For example; einstein's the theory of relativity, as mathematical, is to based on The
Poincere and Minkowsky mathenatics. Big-bang's mathematical conjecture is also to based on Poincare.
Poincare conjecture, we hope that, will open new horizons in cosmology.
Mehmet Taskan
Static links: “blogroll”
Posted by mehmet
8/29/2006 03:48:09 AM



patrick | weblog | photos | videos | audio | profile
Patrick
Age: 23
Gender: Male
Location: New York City
Weblog: patrick
Friend
Nudge
Message
About Me
I work at Xanga :o
More About Me
Interests
Deborah
Occupation
Software Engimineering
Industry
The INTERNET!
Add to Your Profile
Personal info
A/S/L: Age, Sex, Location
Food
Favorite Cuisine
mexican... I like it hot
Free-text info
Favorite Dishes
filet mignon
Favorite Desserts
chocolatey goodness
Contact Info
AIM
MSN
@hotmail.com
ICQ
Add to Your Profile
Instant messenger
and email ids
Feedback
Friends: 80
Memories: 2
Chatboard: 21
Friends (80)
Janet
Matt
cassie
brittney
Links to
friends on
same host

Learning Labels on Multigraphs

- Blogs, blog links, web links, comments etc. implicitly define a (massive) multigraph
- We focus on problems of learning labels
- Our focus is on properties of the blog author such as **age**



- As with all supervised learning, cannot always trust the training data… apparently some people lie about their age

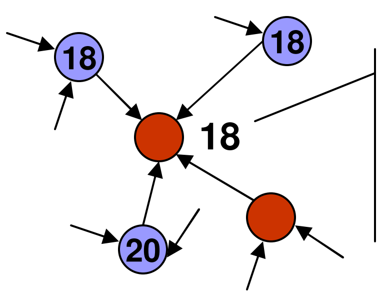
Prior Work on (Multi)graph Learning

- Relational learning: classify objects represented by Relational Database (see work by Getoor et al)
- Typically builds complex models e.g. Relational Markov Networks on relatively small examples (few thousand nodes)
- Our problem is also an instance of semi-supervised learning (input is mix of labelled and unlabeled examples)
- Several works apply matrix decomposition, does not scale well to massive (multi)graphs
- Some work on similar labelling problems on web graph in addition to text (Chakrabarti et al., 1998)

Simple Learning on Graphs

Local: Iterative

Hypothesis: Nodes point to other nodes with similar labels (homophily)

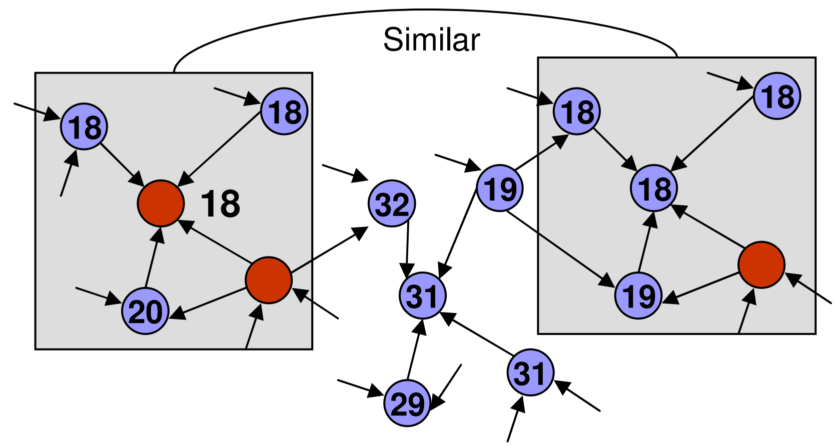


Label is computed from the votes by its neighbors

- Labels are computed iteratively using weighted voting by neighbors

Global: Nearest Neighbor

Hypothesis: Nodes with similar neighborhoods have similar labels (co-citation regularity)

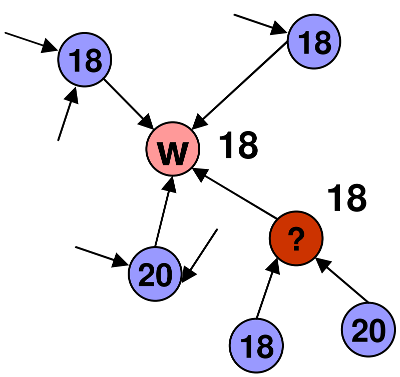


- Label is inferred by searching for similar neighborhoods of labeled nodes

Extend Learning to Multigraphs

Iterative: Pseudo Labels

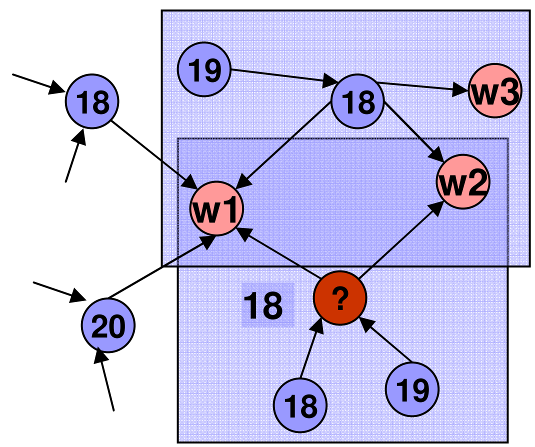
Hypothesis: Web pages link similar communities of bloggers



- Webpages assigned a pseudo label, based on votes by its neighbors

Nearest Neighbor: Set Similarity

Hypothesis: Distance computation is improved with additional features



- Augment distance with similarity between sets of neighboring web-nodes

Implementation Issues

- Preliminary experiments guided choice of settings:
  - Choice of similarity function for NN classifier: used correlation coefficient between vectors of adjacent labels
  - Smoothed feature vector with triangular kernel because of continuity of ages
  - In multigraph case with additional features, extended by blending with Jacard coefficient of set similarity of features
  - Iterative algorithm allocates label based on majority voting
- Experimented with variety of edge combinations: Friends only, blog only, blog+friends, blog+web

Data Collection Summary

Blogger

400K profiles crawled
50K (12.5%) labeled

41K blog nodes
190K blog links
331K web nodes
997K web links

Median: 4 blog links
Median: 3 web links

Most popular weblinks
1. news.google.com
2. picasa.google.com
3. en.wikipedia.org
4. www.flickr.com
5. www.statcounter.com

LiveJournal

300K profiles crawled
124K (41%) labeled

200K blog nodes
404K blog links
289K web nodes
1089K web links

Median: 2 blog links
Median: 4 web links

Most popular weblinks
1. maps.google.com
2. www.myspace.com
3. photobucket.com
4. www.youtube.com
5. quizilla.com

xanga.com

780K profiles crawled
500K (64%) labeled

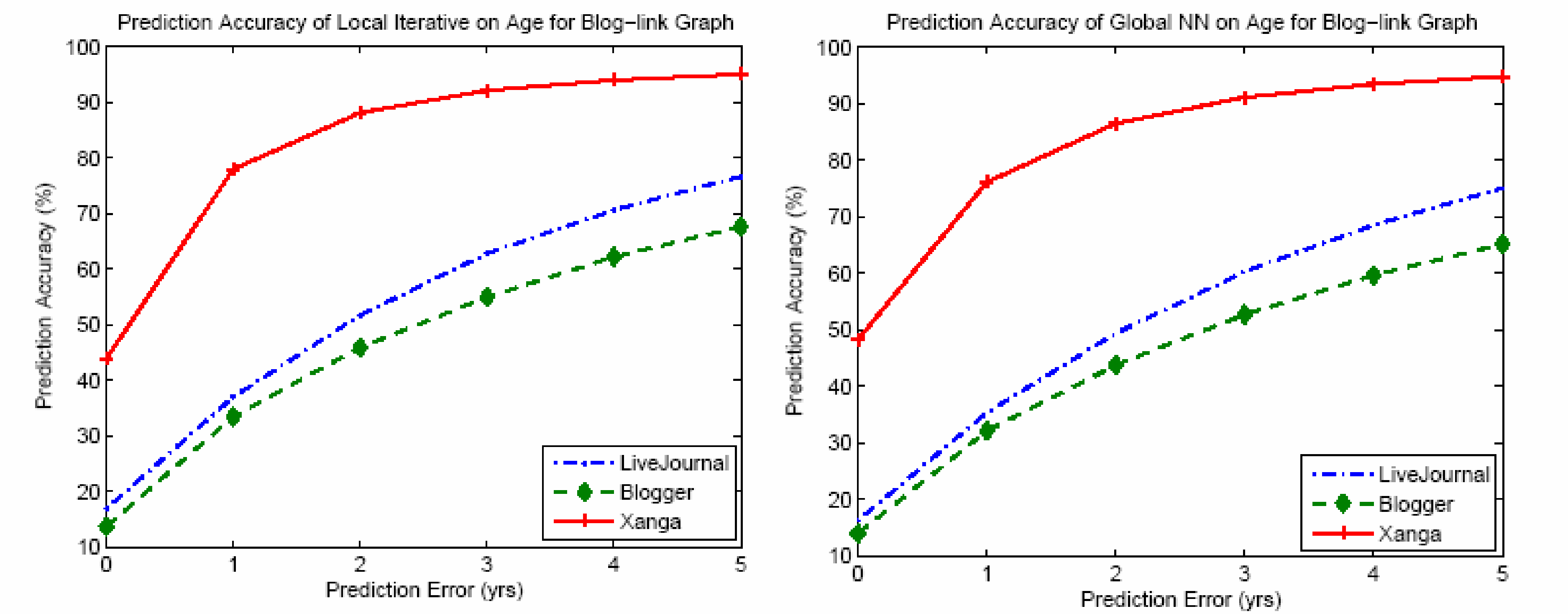
535K blog nodes
3000K blog links
74K web nodes
895K web links

Median: 5 blog links
Median: 2 web links

Most popular weblinks
1. members.msn.com
2. wwp.icq.com
3. edit.yahoo.com
4. www.gottem.net
5. www.crazyarcades.com

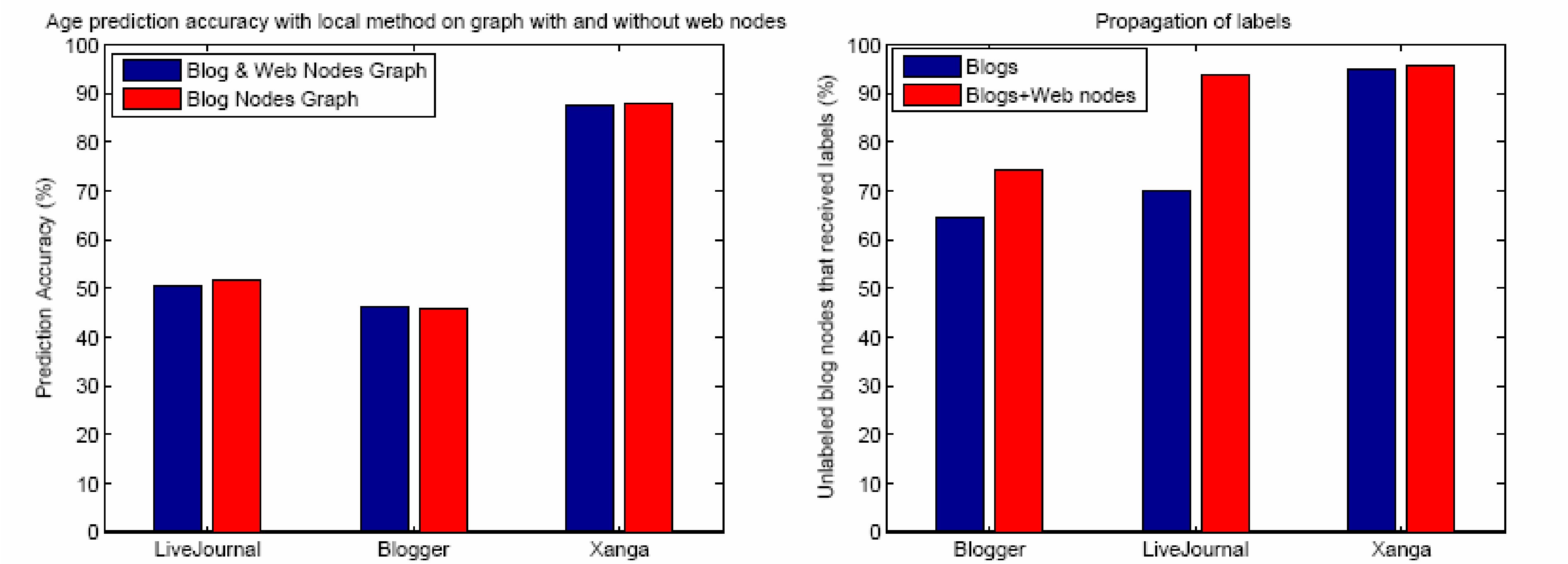
≈50GB of data collected

Accuracy on Age Label



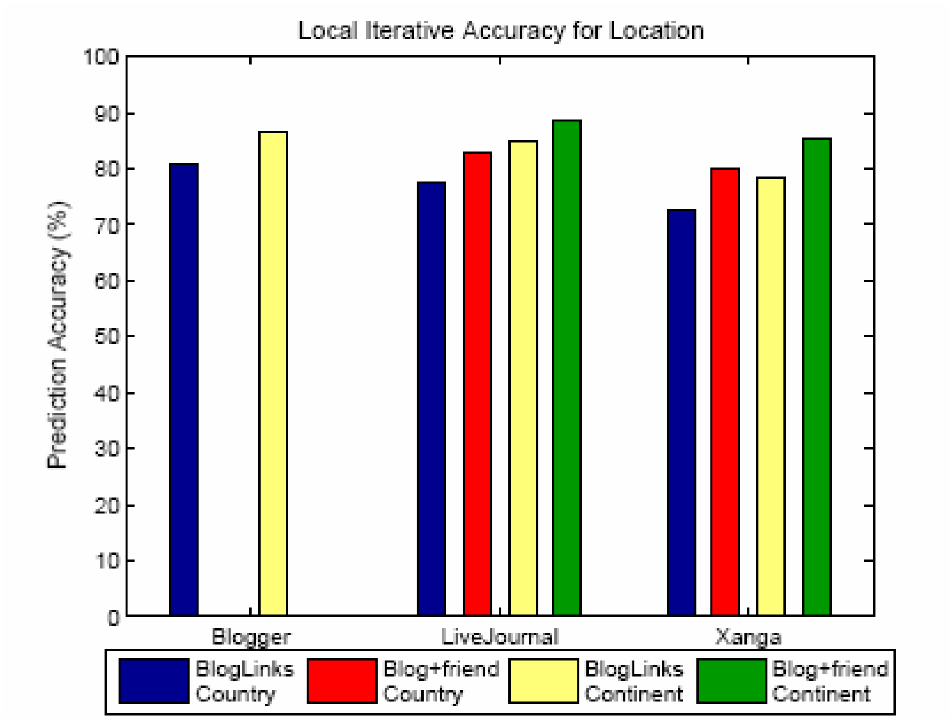
- Similar results on age for both methods, some data sets are “easier” than others, due to density and connectivity
- Local algorithm takes few seconds to assign labels, NN takes tens of minutes (due to exhaustive comparisons)

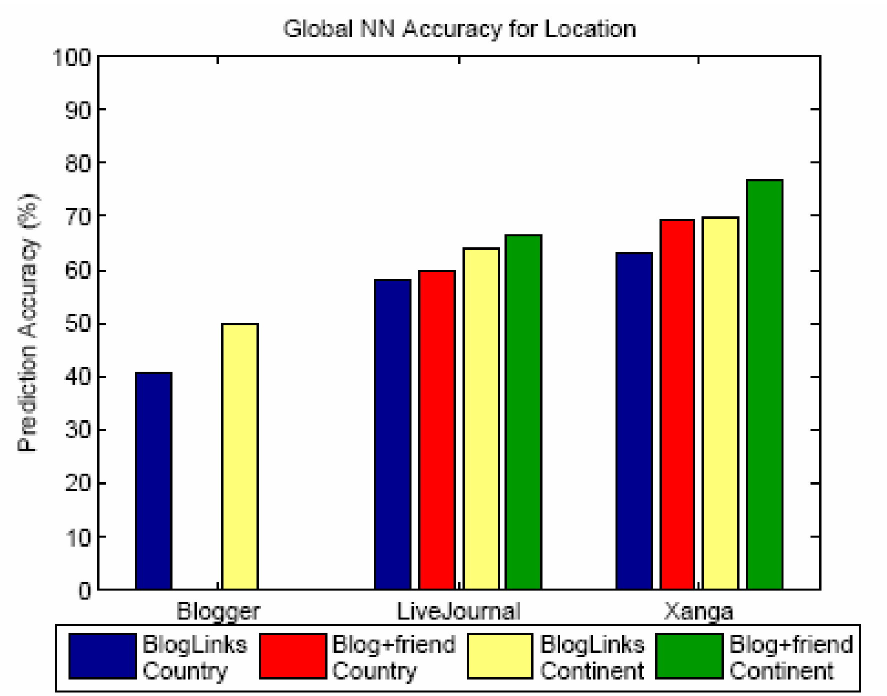
Multigraph Labeling for Age



- Adding web links and using pseudo labels does not significantly change accuracy, but increases coverage
- Assigned age reflects webpage, e.g. bands slipknot (17) vs. Radiohead (28), but also demographics of blog network

Learning Location Labels





- Local algorithm predicts country and continent with high (80%+) accuracy over all data sets, validating hypothesis
- Errors come from over-representing common labels: N. America has high recall, low precision, Africa vice-versa.

Conclusions

- Analyzed performance of simple classifiers for blog data using link and label information only
  - Richness of setting leads to many details: choice of distance, smoothing and voting functions, etc.
  - Links alone still hold a lot of information: 80% accuracy, better than naïve use of standard classifiers
- Simple models are quite limited, do not extend easily
  - Work better for some labels, rely on hypotheses
  - Open to apply and scale richer models (Relational Markov Networks) to blogs
- Need to understand benefit of additional attributes
  - in our expts, extra features did not seem to help