Mining Social Topologies from Email for Online Data Sharing

Diana MacLean, Sudheendra Hangal, Seng Keat Teh,
Monica Lam and Jeffrey Heer

Stanford University

Introduction

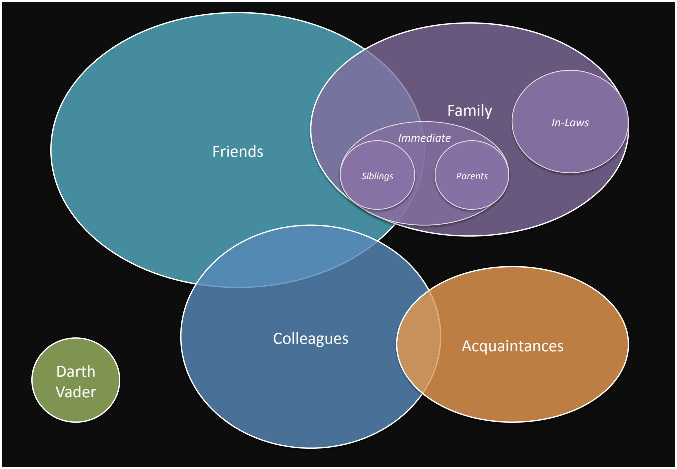
Algorithm

Social Flows Interface

Evaluation

Conclusions

A Social Topology



Social Topologies

The Definition:
"the structure and content of a person's social affiliations, consisting of a set of overlapping social groups and the subset/superset relationships between them."

The important bits:

- overlapping
- nested
- extremely granular

A Snapshot of Today

- Online experience increasingly *social*.
- Social ties handled at *each* point of contact.
    - Facebook friends' lists
    - Gmail contact groups
    - Dropbox communities ...

So what's wrong with this picture?

- Requires *manual* setup (at each point of contact)
- Requires *manual* maintenance (at each point of contact)
- **Does not scale**

Our Vision

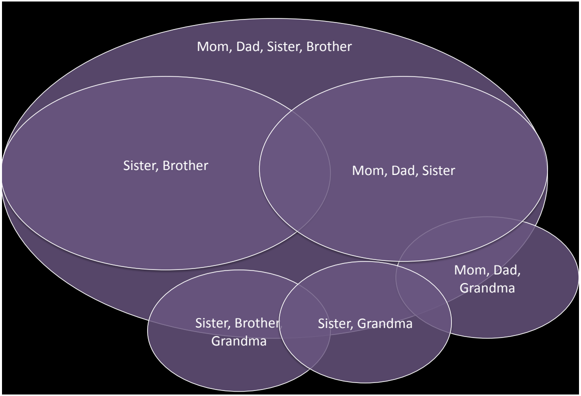
Your Social Topology is captured *latently* in your daily communication patterns anyway (think: e-mail).

- Mine it
- Maintain it
- Port it (or parts of it) to online services

Why e-mail?

- Everyone has it
- Spans several years
- Good labeled data
- Reflects changes in relationships over time

Naive Attempt



Let every unique recipient set be a group.

- too many groups
- incomplete
- lacking macrostructure
- **it's all about pruning**

Our Attempt: Properties

creates a social topology from a sent mail folder that is

- ∼80% smaller than naive model
- ∼95% smaller than naive model with collapsed hierarchy
- easily navigable
- accurate and complete (more about this later)

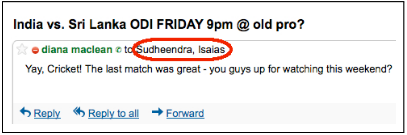
Our Attempt: Outline

4 components:

- Phase 0: Data preparation/cleaning
- Phase 1: Extracting "social molecules"
- Phase 2: Merging "social molecules" into larger groups
- Phase 3: Organizing results into a hierarchy

Phase 1: Extracting "Social Molecules"

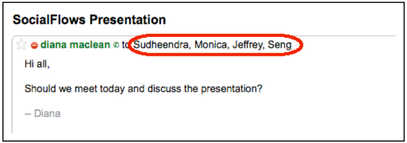
"A social molecule is a small group of people that comprise a relevant, logical social unit according to the users communication patterns."



India vs. Sri Lanka ODI FRIDAY 9pm @ old pro?

diana maclean to Sudheendra, Isaias

Yay, Cricket! The last match was great - you guys up for watching this weekend?



SocialFlows Presentation

diana maclean to Sudheendra, Monica, Jeffrey, Seng

Hi all,

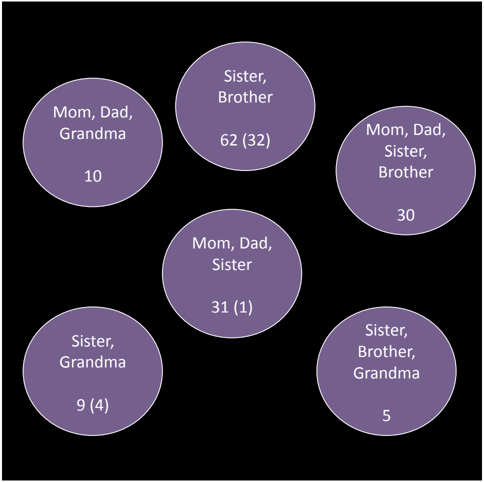
Should we meet today and discuss the presentation?

-- Diana

Obvious proxy: unique, frequent recipient sets

Important property: individuals may belong to several social molecules

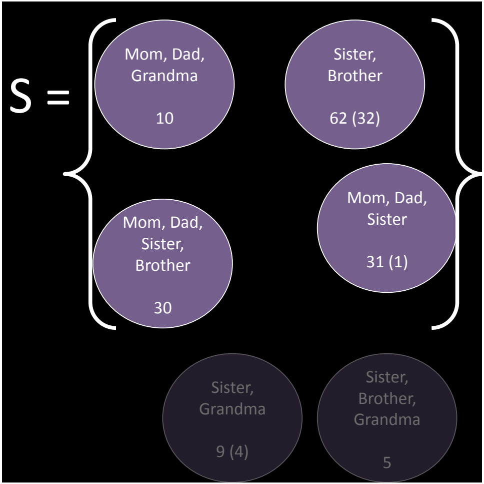
Phase 1: Extracting "Social Molecules"



Let S be the set of social molecules.

Add each unique message recipient set s to S *if* s has high enough frequency in the corpus

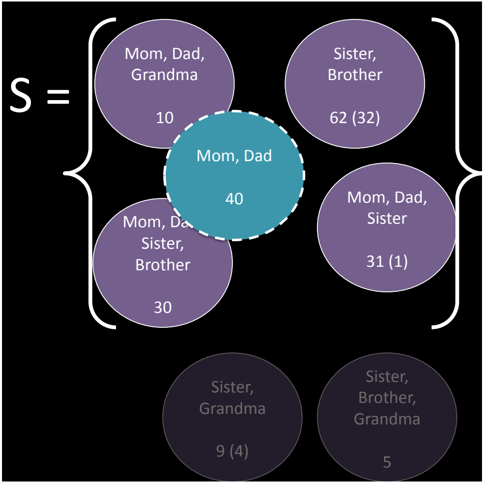
Phase 1: Extracting “Social Molecules”



Let S be the set of social molecules.

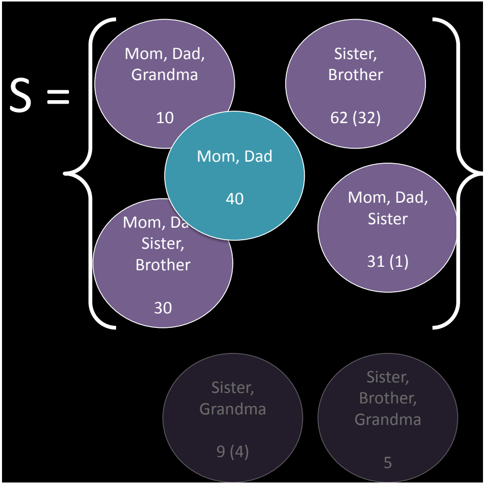
Add each unique message recipient set s to S *if* s has high enough frequency in the corpus

Phase 1: Extracting "Social Molecules"



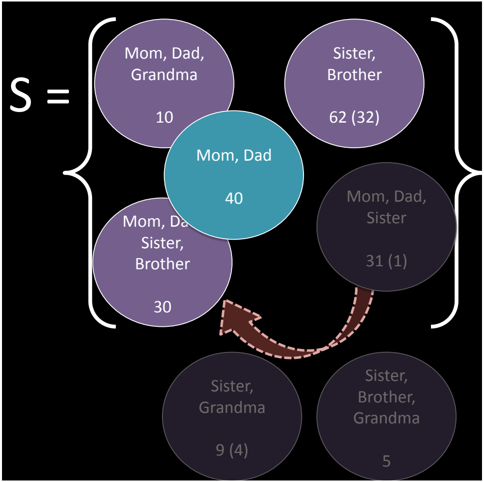
Add to S all the pairwise intersections of S, under the same criteria.

Phase 1: Extracting “Social Molecules”



Add to S all the pairwise intersections of S, under the same criteria.

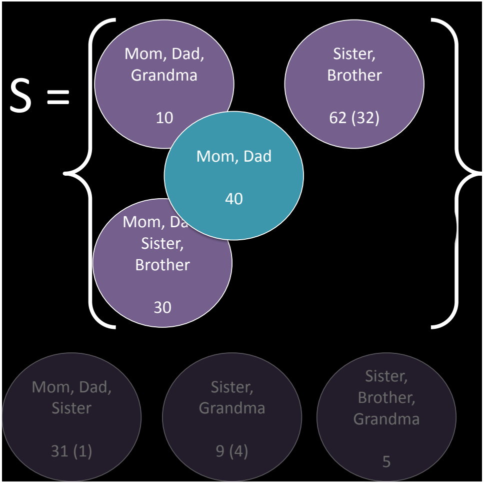
Phase 1: Extracting "Social Molecules"



Retain only those s with sufficient *self identity*

- The *sharing error* between s_i and s_j, $s_j \subset s_i$ is a measure of information leaked if s_i and s_j were merged.
- $serr(s_i, s_j) = \frac{(|s_i| - |s_j|) x (msgs(s_j) - msgs(s_i))}{|s_i| x msgs(s_j)}$
- If $serr(s_i, s_j)$ is large, we retain s_j, otherwise we say s_i *subsumes* s_j

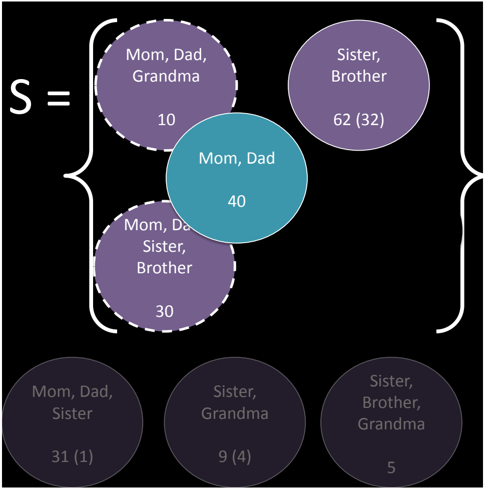
Phase 1: Extracting "Social Molecules"



Retain only those s with sufficient *self identity*

- The *sharing error* between s_i and s_j, $s_j \subset s_i$ is a measure of information leaked if s_i and s_j were merged.
- $serr(s_i, s_j) = \frac{(|s_i| - |s_j|)x(msgs(s_j) - msgs(s_i))}{|s_i|xmsgs(s_j)}$
- If $serr(s_i, s_j)$ is large, we retain s_j, otherwise we say s_i *subsumes* s_j

Phase 2: Merging Social Molecules

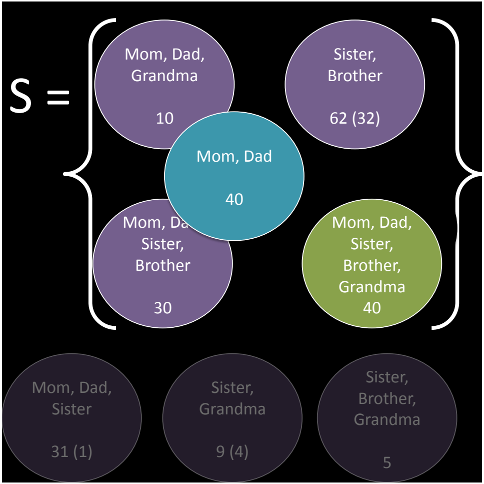


But we're still missing macro-structure.

For each s_i, s_j, add $(s_i \cup s_j)$ to S if s_i and s_j are sufficiently "similar". Note that:

- use Jaccard similarity according to set membership with some threshold
- note: no sets are discarded

Phase 2: Merging Social Molecules

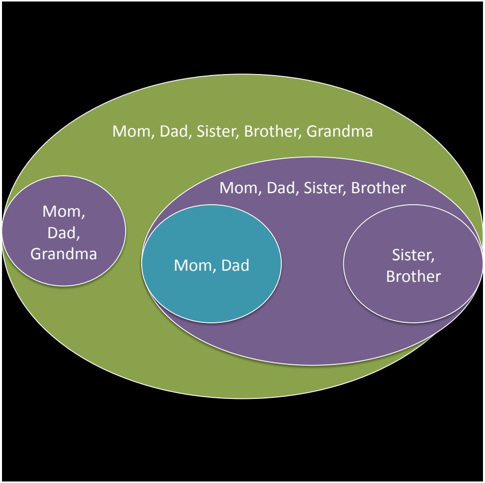


But we're still missing macro-structure.

For each s_i, s_j, add $(s_i \cup s_j)$ to S if s_i and s_j are sufficiently "similar". Note that:

- use Jaccard similarity according to set membership with some threshold
- note: no sets are discarded

Phase 3: Organizing S into a Hierarchy

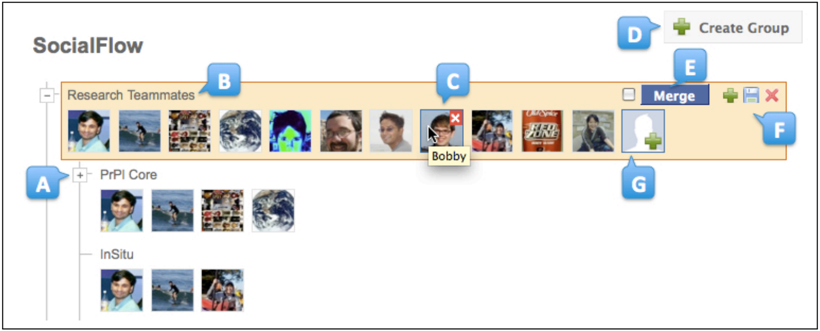


- Find s', the s with greatest *group mass* in S. s' is a parent group
  - *group mass* of s = sum of msgs attributed to each person in s
- Assign all s similar to s' as children groups
  - Again, use Jaccard similarity according to set membership with some threshold

Applications to other data...

Applies to almost any data with group tagging, and in which frequency $\sim$ importance.

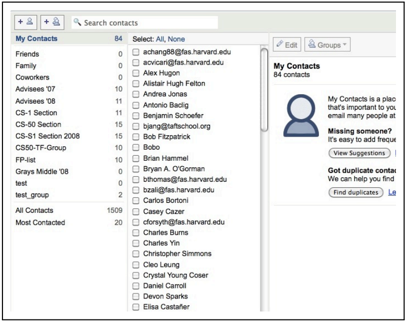
Our interface allows easy browsing and manipulation of social topologies.



Annotated points of interest highlight: (a) hierarchical nesting of subsets; (b) editable group labels; (c) tooltip and delete option on mouse hover; (d) new group creation; (e) group merge tools; (f) additional group editing tools and (g) option to add a new contact.

User Study Evaluation

Using a.) Gmail Contacts tool, and b.) Social Flows, ask users to create partial social topologies congruent to contrived scenarios



My Contacts	84
Friends	0
Family	0
Coworkers	0
Advisees '07	10
Advisees '08	11
CS-1 Section	11
CS-50 Section	15
CS-S1 Section 2008	15
CS50-TF-Group	10
FP-list	10
Grays Middle '08	0
test	0
test_group	2
All Contacts	1509
Most Contacted	20

Select: All, None

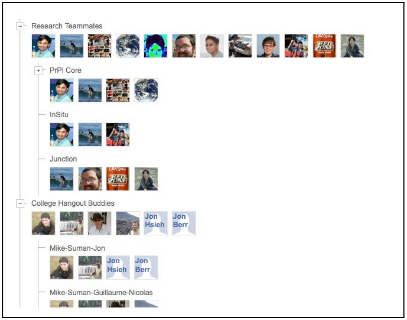
- ☐ achang88@fas.harvard.edu
- ☐ acvicari@fas.harvard.edu
- ☐ Alex Hugon
- ☐ Alistair Hugh Felton
- ☐ Andrea Jonas
- ☐ Antonio Baclig
- ☐ Benjamin Schoefer
- ☐ bjang@taftschool.org
- ☐ Bob Fitzpatrick
- ☐ Bobo
- ☐ Brian Hammel
- ☐ Bryan A. O'Gorman
- ☐ bthomas@fas.harvard.edu
- ☐ bzali@fas.harvard.edu
- ☐ Carlos Bortoni
- ☐ Casey Cazer
- ☐ cforsyth@fas.harvard.edu
- ☐ Charles Burns
- ☐ Charles Yin
- ☐ Christopher Simmons
- ☐ Cleo Leung
- ☐ Crystal Young Coser
- ☐ Daniel Carroll
- ☐ Devon Sparks
- ☐ Elisa Castañer

My Contacts
84 contacts

My Contacts is a plac
that's important to you
email many people at

Missing someone?
It's easy to add freque

Got duplicate contac
We can help you find



User Study Evaluation: Results

6 out of 19 users found the Gmail interface intolerable and quit the task!

User Study Evaluation: Results

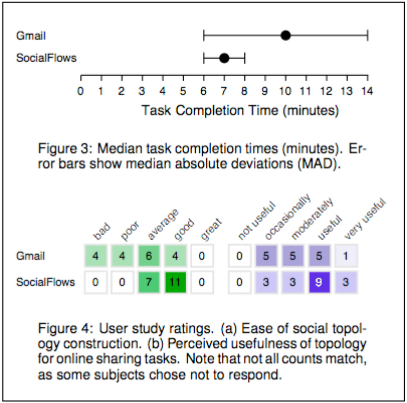


Figure 3: Median task completion times (minutes). Error bars show median absolute deviations (MAD).

	bad	poor	average	good	great	not useful	occasionally	moderately	useful	very useful
Gmail	4	4	6	4	0	0	5	5	5	1
SocialFlows	0	0	7	11	0	0	3	3	9	3

Figure 4: User study ratings. (a) Ease of social topology construction. (b) Perceived usefulness of topology for online sharing tasks. Note that not all counts match, as some subjects chose not to respond.

- Topology creation in Social Flows is significantly *faster*
- Social Flows interface is significantly *easier to use*
- Resulting topologies in Social Flows are significantly *more satisfactory*
- Resulting topologies in Social Flows are *more useful* for online sharing tasks.

User Study Evaluation: Conclusions

- Our algorithmically generated templates:
  - reduce overhead construction time of social topologies
  - reduce cognitive recall required to remember group membership
  - are reasonably accurate
- Our Social Flows interface is an improvement over state of the art in managing social contacts

Take Aways

- Social Topologies data structure
- Algorithm to generate overlapping social groups (first that we know of)
- Port to sites for centralized sharing/access control

Introduction

Algorithm

Social Flows Interface

Evaluation

Conclusions