Web Mining for Implicit User Affinities in On-line Communities

Matthew Smith
smitty@byu.edu
Brigham Young University
March 2005

Overview

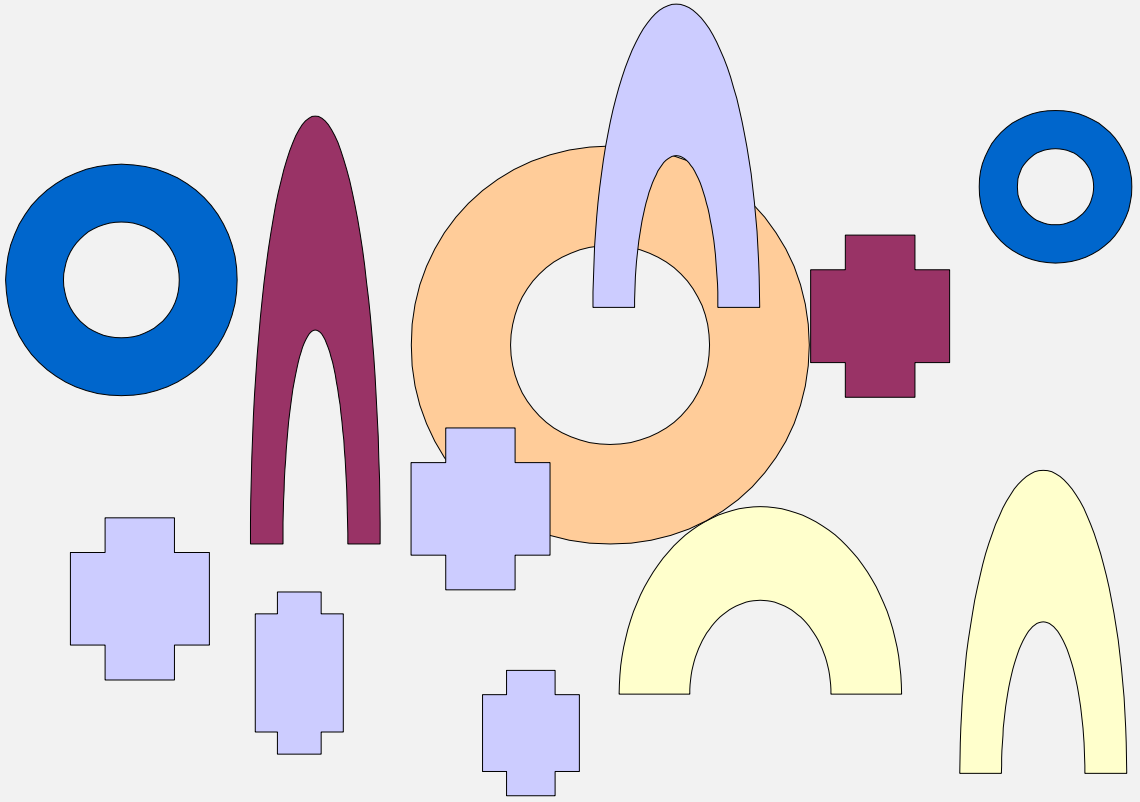
- Introduction
  - Affinity, Hypothesis, Scenario
- User Data
- Identifying Implicit User Affinities
- Quantifying Implicit User Affinities
- Conclusion

Affinity

- Definition:
  - An inherent similarity between persons or things
- Synonyms:
  - Relationship, connection, closeness, association, etc.

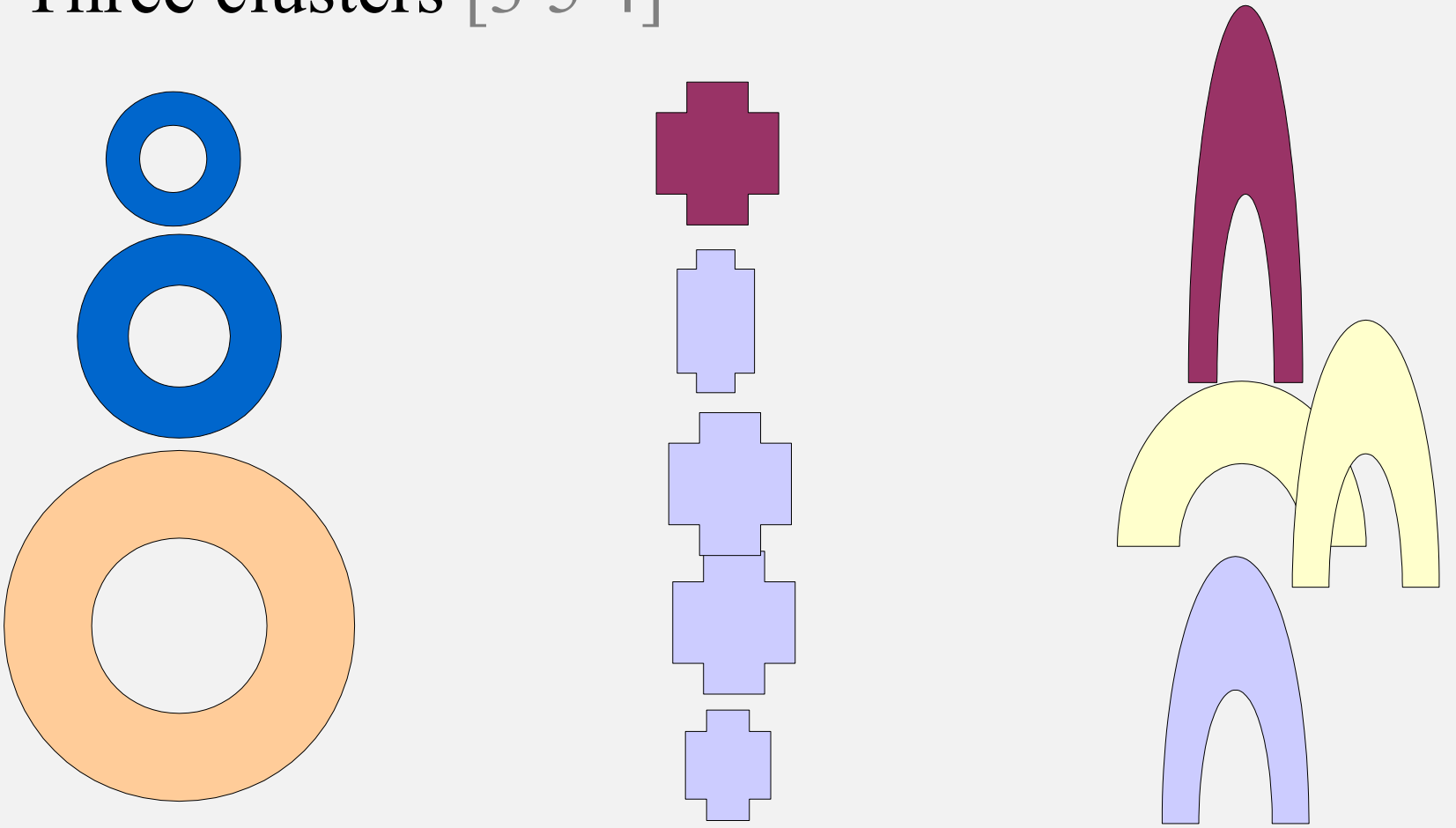
Basic Shapes

- What are some affinities?



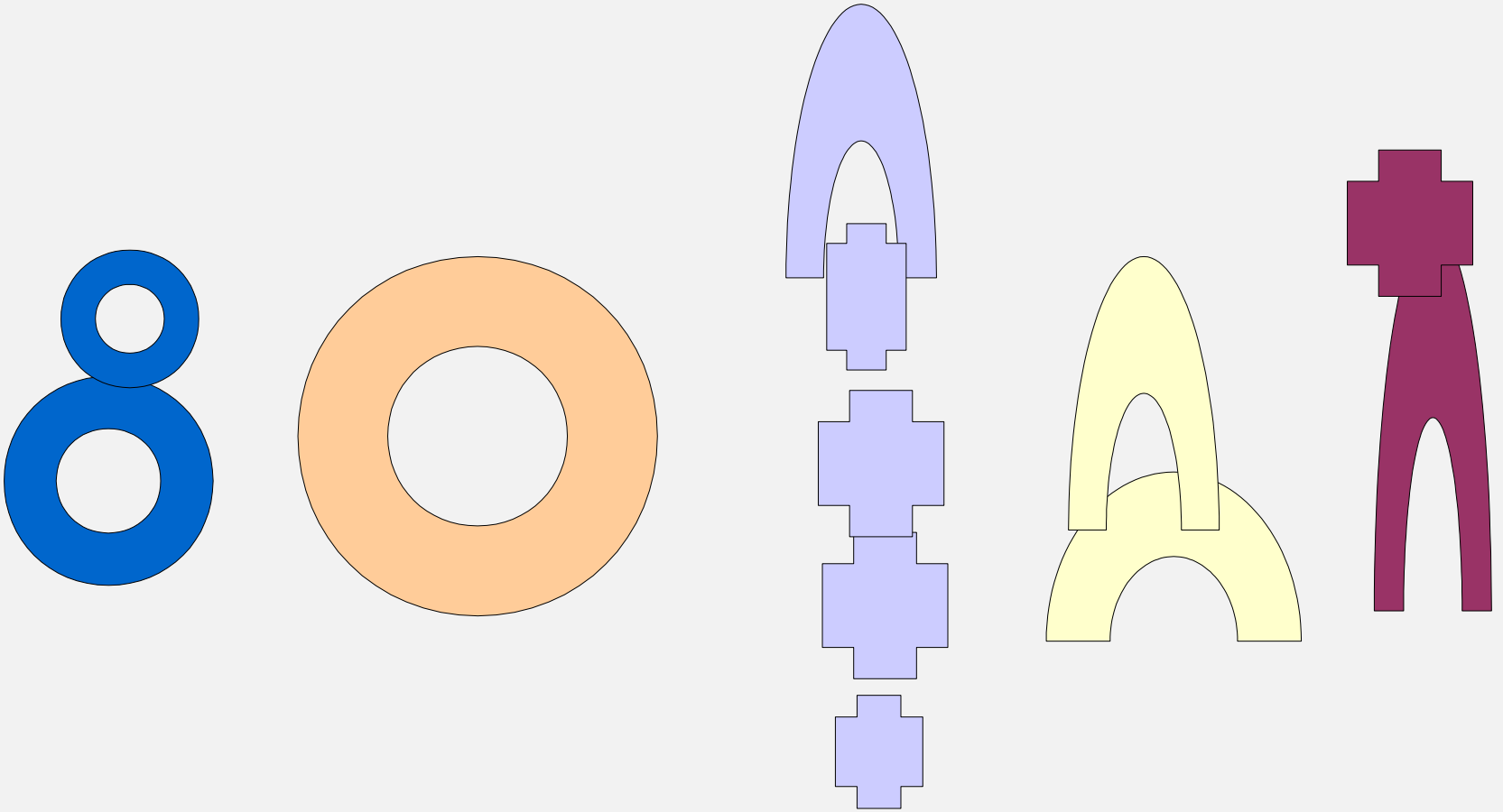
Clustered by Shape

- Three clusters [3 5 4]



Clustered by Color

- Five clusters [2 1 5 2 2]



Hypothesis

- Hypothesis:
  - Simple affinities can be discovered using basic user data within on-line communities
- Why might this be useful?
  - Better integrate new users into on-line communities
  - Interesting to learn how people within a community are related
  - Predict who users might like to meet

Use Case: Scenario

- You join an on-line community
  - *Pick one of your choice:* Google groups, Yahoo! Groups, **LDSMissions.com**, ACM, LDSMingle.com, Orkut, UUG, DevHood, uphpu, kddnuggets, or any other community you can think of.
- You want to know if you have any connections with anyone?
  - Who are those that you might know through other channels?
  - Who in the community do you share commonality (i.e., similar hobbies, similar geography, similar something, etc.) and who might you be interested in meeting?
  - Who of those similar to you is "well-connected" with other individuals you might like to meet?

User Data Collection

- Initial study: www.LDSMissions.com
  - Focus was on the basic data already being collected:

    username, first name, last name, email address, city, state, zip, country, home ward & stake, and mission information (such as city, state/location, zip, country, areas served, companions, mission presidents, etc.)

  - Registration form http://www.ldsmissions.com/us/?action=missionary.register
  - What affinities might this data support?

LDSMissions Data

- Sample:

unique identifier | derived attribute | text mining required | from mission table

username	last name	first name	city	state	country	domain	mission presidents	areas served	companions	mission	mission country	mission location	start date	end date
			WJ	UT	UNITED STATE	msn.com	Pres Andrew Day	Taubate' , Soracaba	Sis. Silva, Sis Nali	Brazil Sao Paulo North	Brazil	Brazil	Jul 1988	Jan 1990
			MUNTINLUPA	CITY	PHILIPPINES	YAHOO.COM	PRESIDENT CAR	CATBALOGAN CITY	NGAWAKA,VALEN	Philippines Tacloban	Philippines	Philippines	May 1994	Mar 1996
			New Hartford	NY	UNITED STATE	email.byu.edu	Kradolfer	Alamogordo, Maran	Wardell, Staniszev	Arizona Tucson	United States	Arizona	Oct 2002	Apr 2004
			Elford	BC	CANADA	hotmail.com	Cook, Fillmore	North Philly, Pottsto	Bartlett, Nukaya, L	Pennsylvania Philadelphi	United States	Pennsylvania	Mar 2003	Mar 2005
			Orem	UT	UNITED STATE	hotmail.com	President Garrett,	Crescent View Ward	Elder Moore, Elde	Canada Calgary	Canada	Canada	Feb 2003	Mar 2005
			Bury,Lancashir		UNITED KINGD	yahoo.co.uk	Pres.Eldon.N.Tan	isle of white,salisbu	15.companions in	England London South	England	United Kingd	Nov 1983	Sep 1985
			Pleasant Grove	UT	UNITED STATE	yahoo.com	Dewitt, Goodman,	Cosenza, Siracuza,	Ellsworth, Greenh	Italy Catania	Italy	Italy	Nov 1994	Nov 1996
			Papatoetoe		NEW ZEALAND	hotmail.com	Pres.Wellis	Honolulu- Hawaiika	Sis.Nishime,Sis.B	Hawaii Honolulu	United States	Hawaii	Sep 2002	Mar 2004
			Fairfax	VA	UNITED STATE	cox.net	Gary E. O'Brien	Biel, Ebnat Kappel,	Alan McLean, Mich	Switzerland Zurich	Switzerland	Switzerland	Nov 1974	Nov 1976
			Ogden	UT	UNITED STATE	yahoo.com	Christiensen	Pdte. Roque Ensan	Williams, Enderle	Argentina Resistencia	Argentina	Argentina	Jan 1999	Jul 1999
			mesa	AZ	UNITED STATE	yahoo.com	ronald g. davis	madrid barrios:7,9,1	lots	Spain Madrid	Spain	Spain	May 2003	May 2005
			Rexburg	ID	UNITED STATE	hotmail.com	President Hanks,	Irving sign, Irving Sp	Hermana Lyddon,	Texas Dallas	United States	Texas	Jun 1998	Dec 1999
			DeWinton	AB	CANADA	hotmail.com	Frank Bradshaw	Escindido, Carlsbad	, Miramar, Cardiff	California San Diego	United States	California	Jan 1973	Dec 1974
			Provo	UT	UNITED STATE	comcast.net	Hickman, Christia	Southport, Clinton, F	Golden, Loosli, W	North Carolina Raleigh	United States	North Carolir	Mar 1993	Feb 1995
			Omaha	NE	UNITED STATE	yahoo.com	H. Ray Hart	Metz, Boulogne sur	Mer, Cambrai, Brus	Belgium Brussels	Belgium	Belgium	Jun 1997	Jun 1999
			Bountiful	UT	UNITED STATE	mac.com	Sorensen, Thacke	Urbana,Springfield,	Johnson,Ohlson,J	Illinois Peoria	United States	Illinois	Jun 1997	Nov 1998
			Salt Lake City	UT	UNITED STATE	yahoo.com	Williams	La Roche sur Yon, I	Young, Hill, Grace	France Bordeaux	France	France	Jan 1999	Dec 2000
			salt lake city	UT	UNITED STATE	yahoo.com	Rod Tueller, Max S	Ponchatoula, New C	Millar, Holt, Richar	Louisiana Baton Rouge	United States	Louisiana	Jan 1999	Dec 2000
			salt lake city	UT	UNITED STATE	yahoo.com	Parkers, Morenos	Las Delicias, Tarija,	Matsumura(steph	Bolivia Cochabamba	Bolivia	Bolivia	Jan 1999	Jul 2000
			antioch	CA	UNITED STATE	aol.com				Venezuela Valencia	Venezuela	Venezuela	Apr 1999	Mar 2001
			Kaysville	UT	UNITED STATE	msn.com	Schreiber		Watts, Schmidt, Pt	Germany Hamburg	Germany	Germany	Jan 1981	Aug 1982
			The Woodlands	TX	UNITED STATE	vogtengineering	J. Ballard Washbu	Prescott, Phoenix, T	Suzette Whiting, H	Arizona Phoenix	United States	Arizona	Jan 1988	Aug 1989
			Taylorsville	UT	UNITED STATE	hotmail.com	President Bernard	Somerset, Georgetc	Pann, Christensen	Kentucky Louisville	United States	Kentucky	Dec 2002	Dec 2004
			Norman	OK	UNITED STATE	hotmail.com	Marshal, Cahoon	Veracruz, Puebla, Al	Northcutt, Borget,	Mexico Veracruz	Mexico	Mexico	Apr 1979	Apr 1981
			CANYON COUI	CA	UNITED STATE	YAHOO.COM	PRESIDENT LEE	HONDURAS TEGUC	IGALPA	Honduras Tegucigalpa	Honduras	Honduras	Jan 2004	Aug 2005
			Cedar Rapids	IA	UNITED STATE	mchsi.com	Gosta Berling, Joh	Bergen, Sarpsborg,	Gary Stoddard, Ste	Norway Oslo	Norway	Norway	Jan 1975	Jan 1977
			Sandy	OR	UNITED STATE	netzero.com	Hobbs, Hunter			New Hampshire Manches	United States	New Hamps	Feb 1997	Feb 1999
			los angeles	CA	PHILIPPINES	yahoo.com.	Daniel rogers	kansas city	sister burgess	Missouri Independence	United States	Missouri	Oct 1998	Sep 2000
			Boise	ID	UNITED STATE	yahoo.com	Donald Hinton, Te	Aberdeen, Ngau Tai	Ure, J. Chan, Vale	China Hong Kong	China	China	Sep 2002	Oct 2004
			Brawley	CA	UNITED STATE	juno.com	Paul Felt	White Cone, Twin L	Phil Smith, E. Rog	Arizona Mesa	United States	Arizona	Jan 1970	Jan 1972

Hidden to preserve anonymity

Identifying Implicit Affinities

- How can affinities be implied?
- Method: **Similarity Clustering and Aggregation**
  - Names: first, last, middle, maiden
  - Mission: name, presidents, companions
  - Geographically
    - Mission: region, country, state/location, city, zip, areas served
    - Home: region, country, state, city, zip, home stake, home ward
  - Email address domains (i.e., hotmail.com, yahoo.com, msn.com, byu.edu, etc.)

Data Aggregation

- Email Domains

domain	count
hotmail.com	4863
yahoo.com	2316
aol.com	1413
juno.com	687
email.byu.edu	500
msn.com	302
earthlink.net	124
cc.usu.edu	115
byu.edu	101
cs.com	97
excite.com	93
usa.net	77
netzero.net	68
netscape.net	67
home.com	65
cox.net	62
byui.edu	62
myldsmail.net	59
worldnet.att.net	59
attbi.com	53
comcast.net	50
latinmail.com	49
byuh.edu	49
uswest.net	46
prodigy.net	44
mstar2.net	41
sisna.com	39
infowest.com	38
weber.edu	35

Data Aggregation

- First names

first_name	count		
David	385	Adam	124
Michael	284	Kevin	121
Ryan	224	Chris	116
Scott	209	Brandon	115
Jason	200	Richard	112
John	181	Paul	111
Daniel	174	Andrew	110
Jared	167	Joshua	110
Matthew	160	Steven	109
Robert	149	Matt	104
Brian	147	Jeremy	104
Mark	146	Christopher	101
James	139	Justin	99
Aaron	132	...	
Nathan	125	Burdette	1
Eric	125		
Benjamin	124		

- Last names

last_name	count		
Smith	168	Larsen	47
Johnson	126	Wilson	45
Jones	104	Wright	45
Anderson	89	Adams	42
Brown	82	Hall	41
Christensen	74	Harris	40
Jensen	74	Hill	39
Hansen	72	White	38
Nelson	67	Walker	36
Peterson	66	Nielsen	33
Taylor	61	Hatch	33
Clark	56	Sorensen	32
Williams	55	Olsen	32
Davis	53	...	
Allen	49	Pixton	1
Thompson	48		
Miller	48		

Data Aggregation

- Geography (Country, State, City)

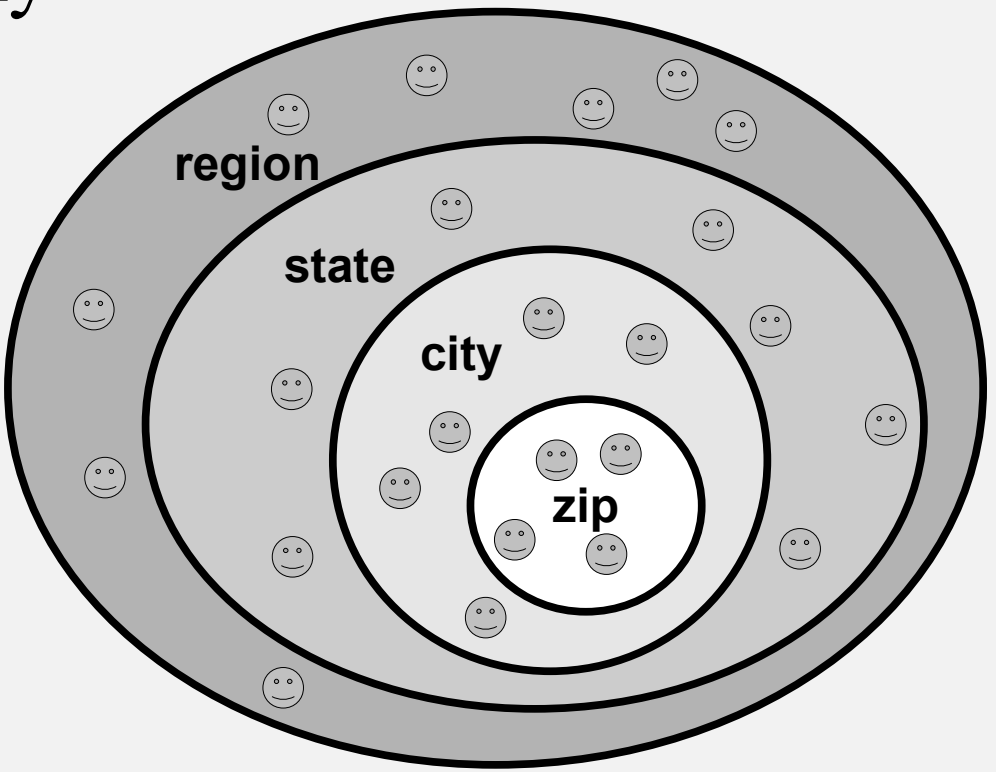
country	count
United States	14272
Canada	320
Philippines	256
Mexico	213
Brazil	155
Chile	111
Australia	110
United Kingdom	110
Argentina	95
New Zealand	79
Peru	69
Spain	33
Guatemala	26
Colombia	23
France	21
Ecuador	21
Bolivia	20
Germany	20
Uruguay	18
Venezuela	18
Sweden	16
Netherlands	15
Puerto Rico (US)	15
Portugal	13
Switzerland	13
Italy	11
Costa Rica	11

state	count
UT	6223
CA	1654
ID	1154
AZ	1116
WA	552
TX	478
NV	316
OR	305
CO	245
HI	209
ALB	193
FL	189
VA	135
GA	122
NM	109
WY	105
MO	92
NC	90
AL	87
OH	87
IL	85
TN	81
IN	76
NY	76
MI	65
BC	64
MD	61
AK	58
KS	57

city	count
Provo	751
Salt Lake City	566
Orem	479
Mesa	402
Sandy	297
Ogden	189
LOGAN	175
Layton	167
Bountiful	165
Las Vegas	164
West Jordan	157
Idaho Falls	146
Rexburg	132
St. George	121
pocatello	111
Taylorsville	109
Boise	107
Gilbert	103
West Valley City	102
SLC	98
Phoenix	98
Murray	98
laie	97
Cedar City	91
South Jordan	86
Kaysville	86
Farmington	84
Draper	77
Springville	72

Visualization: Venn Diagram
(stacked)

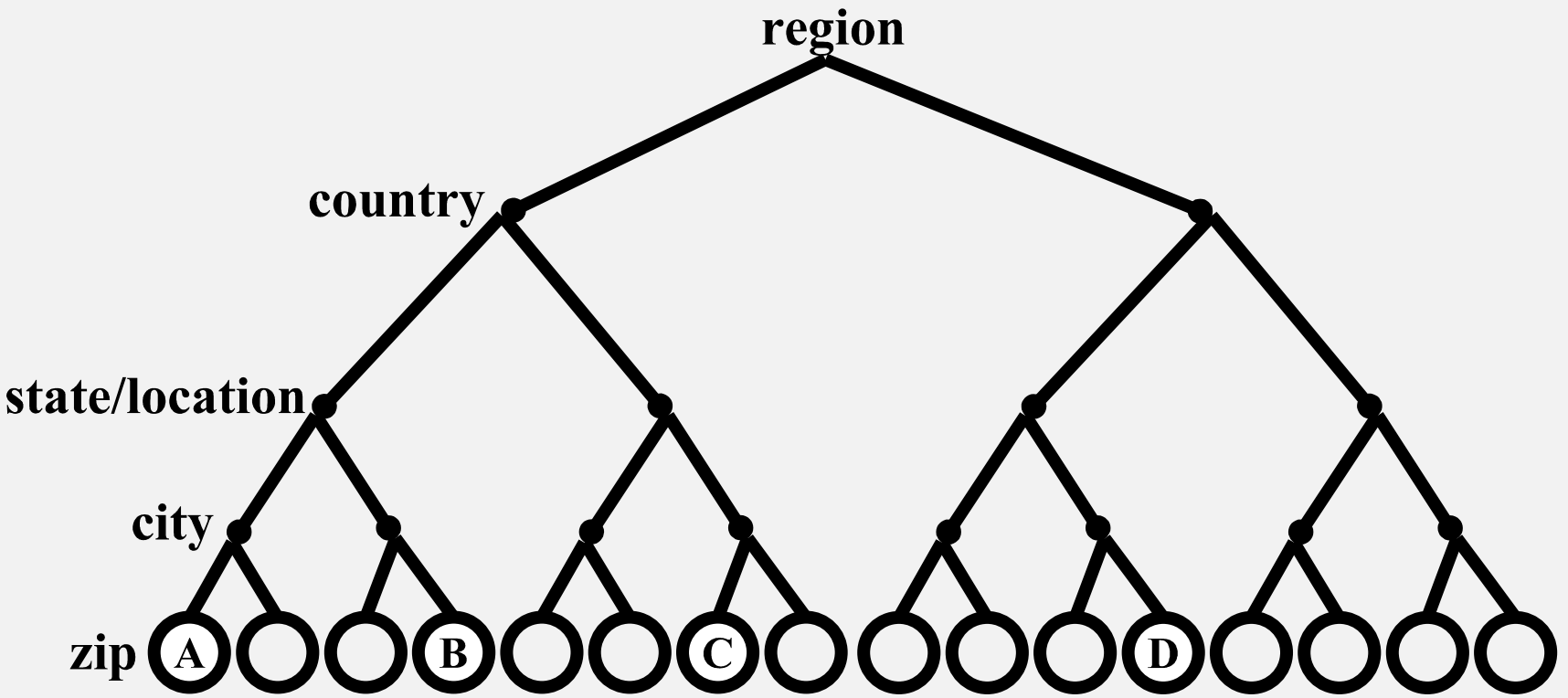
- Geography



Which group is most interesting?

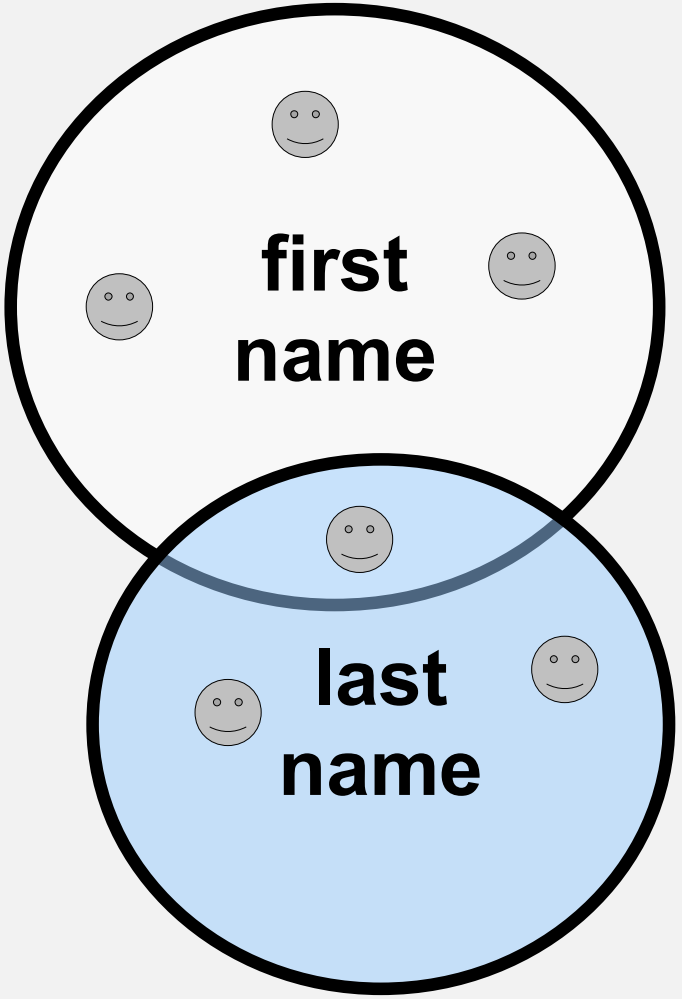
Visualization: Attribute Hierarchy

- Geographical (related attributes)



Visualization: Venn Diagrams

- Names
  - first name match
  - last name match
  - both match



LDSMissions Example

- Login to website...
  - http://www.ldsmissions.com

Quantifying Implicit Affinities

- Why? Important for comparing affinities
- All affinities are not equal
- Affinities are valued differently among users
- Affinities can be aggregated to form a combined affinity score
  - In some cases, quantification will be most useful as a combined calculation

Approach

- Weight affinities equally
  - easiest to implement but less accurate
- Survey all users, then weight affinities accordingly
  - Result may be applied either:
    - Individually
      - Customized affinity scores for each user
    - Collaboratively
      - Average of all user surveys
- Weight affinities by an expert
- Other ideas?

Conclusion

- Useful affinities can be discovered implicitly
  - Using already collected data.
- It is a novel approach to auditing your data
- It is a small world after all.
  - That is, most people are connected through some sort of affinity.
  - All people are are connected through usually small affinity chains.