
Topicmarks

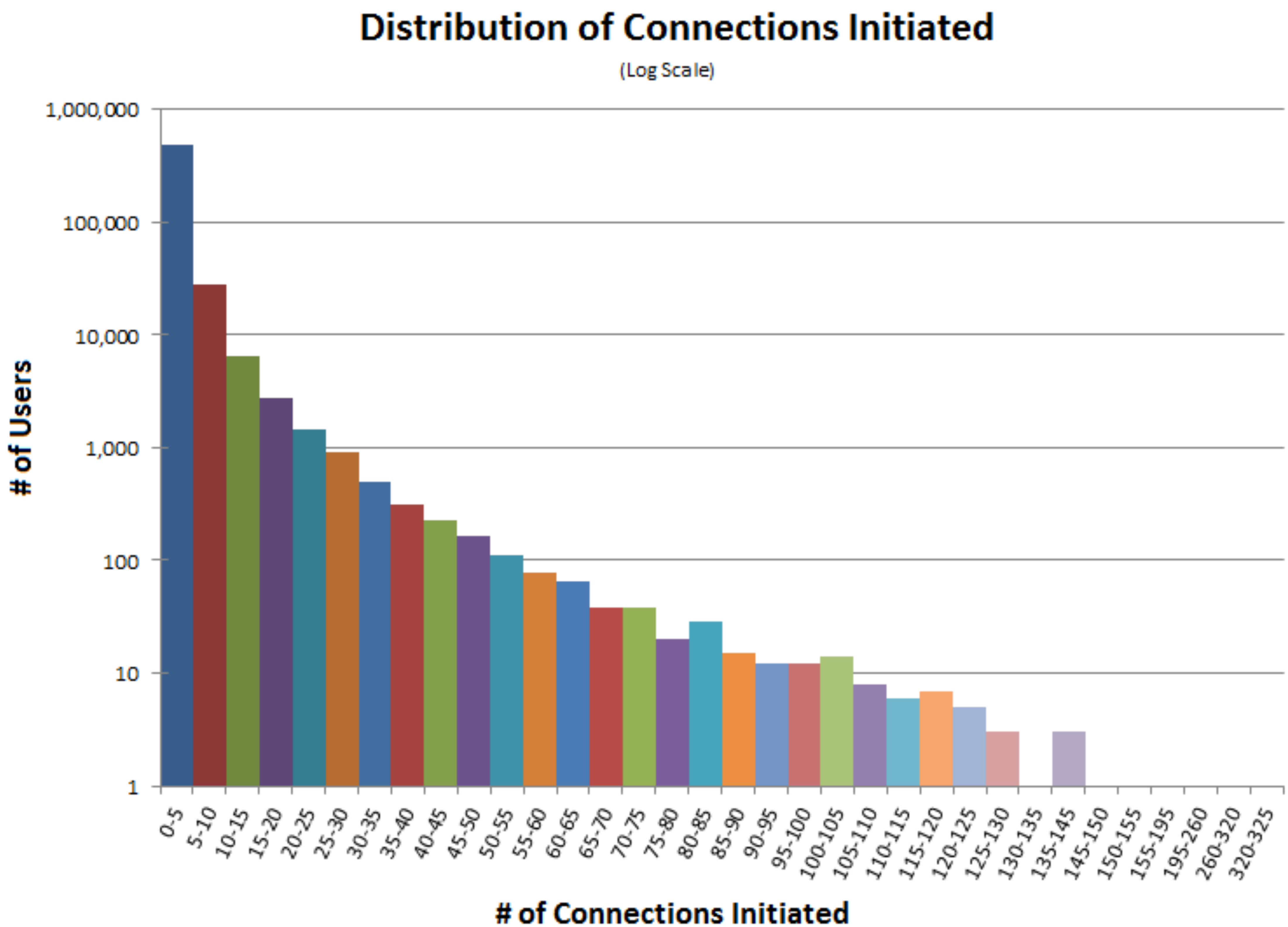
# Overview

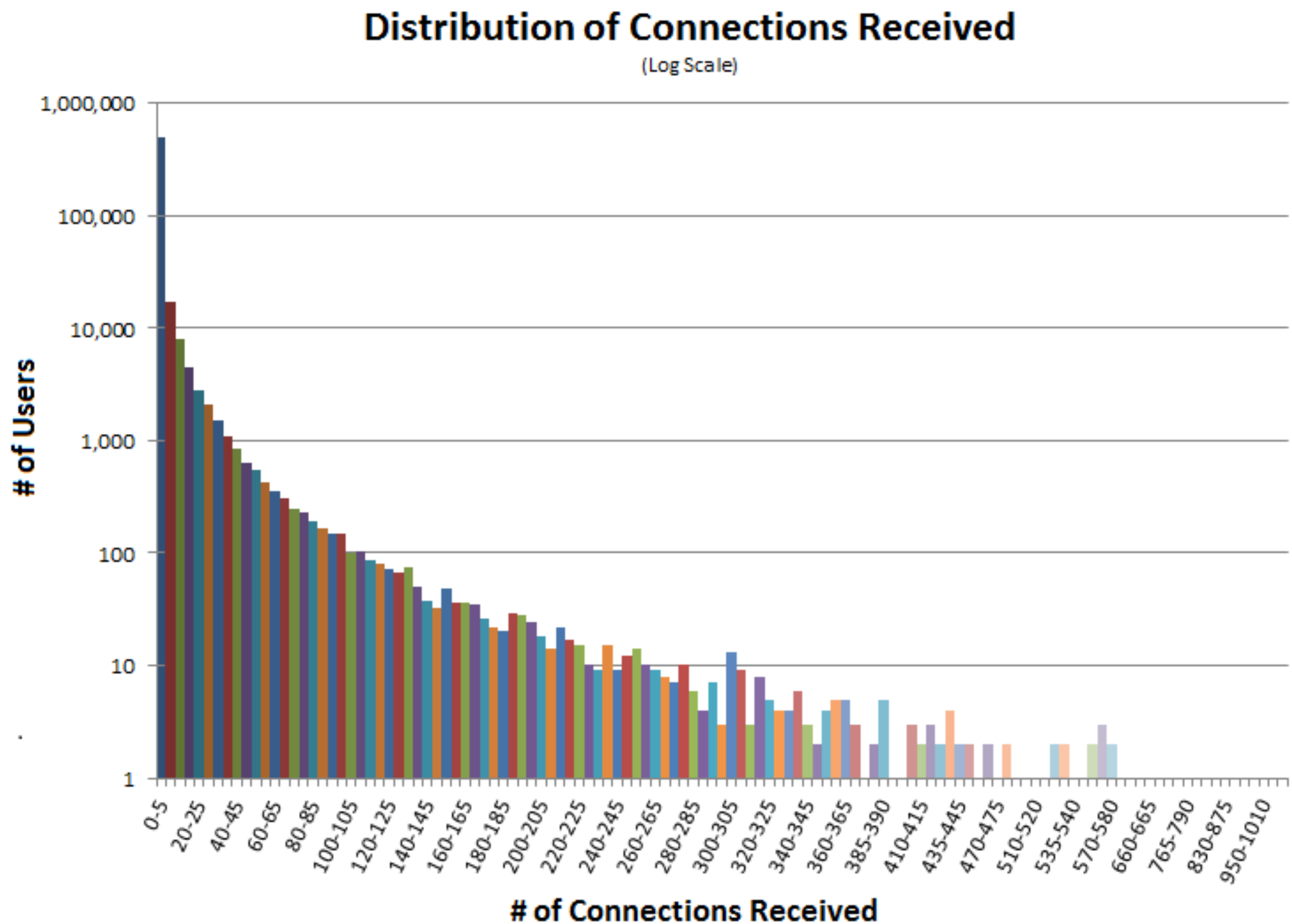
- IM Data Overview
- Working with the Data
- Making Recommendations
- Q & A

# IM Data Overview

# Chat Message Content

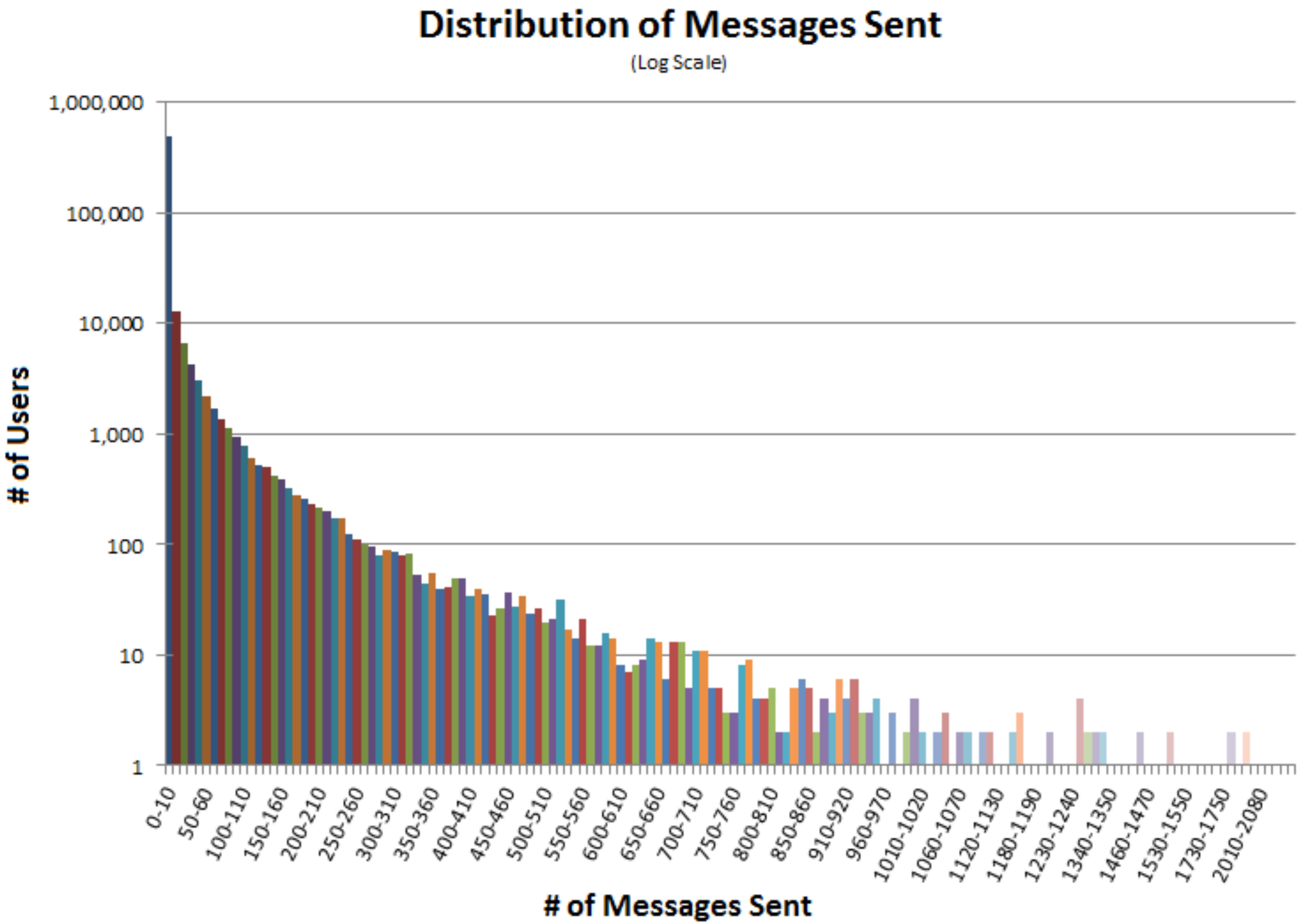
- Content to analyze
    - 2,940,566 messages sent
    - 123,958,816 text characters in messages
    - 13,639,344 terms (excluding stop-words)
- Senders/Receivers
    - 90,779 senders of messages (Californians only)
    - 493,031 recipients of messages
    - 51,752 senders/receivers of messages
- 532,058 profiles based on chat data
- 1,088,099 directional message exchanges


Distribution of Connections Initiated
(Log Scale)
1,000,000
100,000
10,000
1,000
100
10
1
# of Users
0-5
5-10
10-15
15-20
20-25
25-30
30-35
35-40
40-45
45-50
50-55
55-60
60-65
65-70
70-75
75-80
80-85
85-90
90-95
95-100
100-105
105-110
110-115
115-120
120-125
125-130
130-135
135-145
145-150
150-155
155-195
195-260
260-320
320-325
# of Connections Initiated


Distribution of Connections Received
(Log Scale)
# of Users
1,000,000
100,000
10,000
1,000
100
10
1
0-5
20-25
40-45
60-65
80-85
100-105
120-125
140-145
160-165
180-185
200-205
220-225
240-245
260-265
280-285
300-305
320-325
340-345
360-365
385-390
410-415
435-445
470-475
510-520
535-540
570-580
660-665
765-790
830-875
950-1010
# of Connections Received

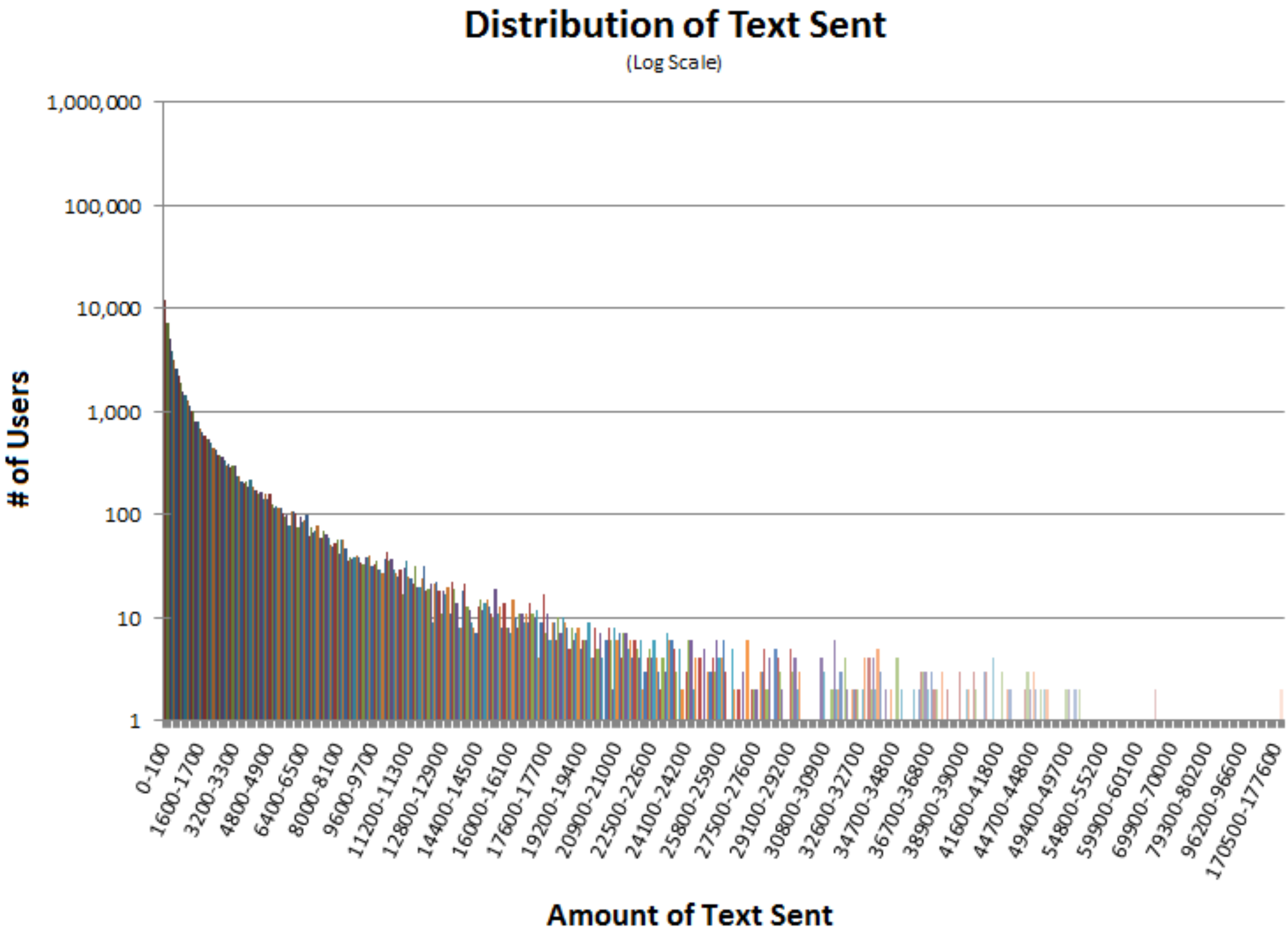


Topicmarks


Distribution of Messages Sent
(Log Scale)
# of Users
1,000,000
100,000
10,000
1,000
100
10
1
0-10
50-60
100-110
150-160
200-210
250-260
300-310
350-360
400-410
450-460
500-510
550-560
600-610
650-660
700-710
750-760
800-810
850-860
910-920
960-970
1010-1020
1060-1070
1120-1130
1180-1190
1230-1240
1340-1350
1460-1470
1530-1550
1730-1750
2010-2080
# of Messages Sent



Topicmarks



Distribution of Text Sent
(Log Scale)
1,000,000
100,000
10,000
1,000
100
10
1
# of Users
0-100
1600-1700
3200-3300
4800-4900
6400-6500
8000-8100
9600-9700
11200-11300
12800-12900
14400-14500
16000-16100
17600-17700
19200-19400
20900-21000
22500-22600
24100-24200
25800-25900
27500-27600
29100-29200
30800-30900
32600-32700
34700-34800
36700-36800
38900-39000
41600-41800
44700-44800
49400-49700
54800-55200
59900-60100
69900-70000
79300-80200
96200-96600
170500-177600
Amount of Text Sent

# Data Challenges

- Lack of grammar, spelling errors, heavy use of vernacular/slang
- Questionable literacy levels for many users (not just chat language)
- Very few salient topics
- Spanglish text
- Inconsistent send/receive patterns - user availability is an unknown
- Data snapshot makes it hard to determine good connections (no previous history)
- No gender identification
- No profile data



Topicmarks

# Data Observations

- Chat messages reflect the pathos of life
  - Everyday life concerns are paramount (work, family, problems)
  - The need for relationships trumps need for sex
- Patterns that emerge
  - Women tend to have more connections received than initiated
  - Common misspelling/vernacular (chillen, shyt, wassup, nuttin, kool)
  - Meetup concerns (distance is an issue)
- Mirroring in message exchanges
  - "You are the sum of your send and receive messages . . . ."
- People are multi-dimensional
  - The occurrence of a single term does not make you a pervert (e.g. "fisting")

# Working with the Data

# Data Classification

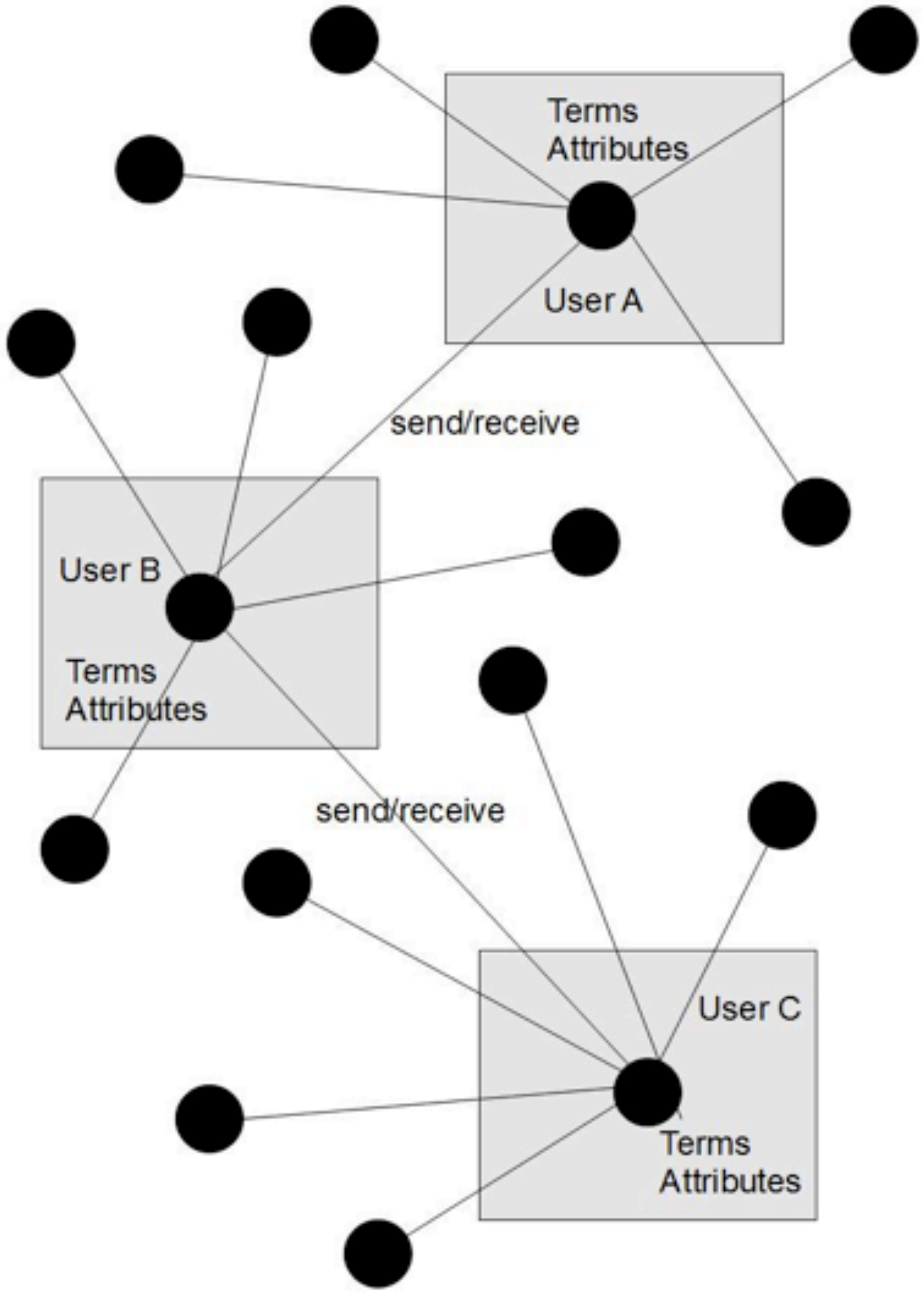
- Existing ontologies/taxonomies of no use
- Created custom classes
- Created corpus lexicon derived with text from all messages (frequency, message count)
- Created classes (called attributes) based on data interpretation
- Created classification sets for wordnet, vernacular/slang and combined (top 1000 terms for each)

# User Attributes

- English - english predominance
- ProperEnglish - proportion of dictionary english versus slang/vernacular
- VocabularyRange - the number of distinct terms in messages
- Raw - use of sexual terms (... you get the idea ...)
- Empathy - the ability to interact (sorry, understand, appreciate, ...)
- Outreach - explicit meetup terminology (sms, email, phone numbers, ...)
- Relationship - blandishments/endearments (sweetie, cutie, boyfriend, girlfriend,...)
- Tagged - tagged community interaction (tagged, pet, buy, add, ...)
- Appearance - physical description (short, tall, hair, face, neck, fat, ...)
- Work - terms related to work (work, job, boss, manager, shift, ...)
- Behavior - activity descriptors (chillin, surfing, watching, sleeping, ...)
- Education - discussion centering around education (school, classes, teacher, professor, ...)
- Family - references to family (mother, father, daughter, son, children, etc.)
- Money - monetary concerns (money, cash, dollars, car, apartment, house, ...)
- Adversity - life's trials and tribulations (divorced, sick, ill, drugs, ...)
- Religiosity - prevalence of religion in discourse (blessed, rosaries, jesus, god, ...)
- Prosperity - life positives (trip, vacation, holidays, army, navy,...)

# User Discourse Analysis

- For each user, construct a discourse set based on send and receive messages
  - note: some people send and receive, some only send, some only receive
- Compute the most important terms for the user based on tf-idf calculations using corpus lexicon
- Based on these terms, identify the prevalent discourse attributes for each user
  - attributes based on WordNet terms
  - attributes based on vernacular/slang
  - attributes based on the combination of WordNet and vernacular
- Construct user profile aspects based on slang, WordNet and total word usage



# Making Recommendations

# IM Analysis

- From the raw data, build a communication graph where message exchange between two tagged members is greater than one
  - 351,459 sender-receiver edges (directed graph)
  - 60,602 senders of messages (67% of raw data senders)
  - 196,794 receivers of messages (40% of raw data receivers)
  - 24,875 senders/receivers of messages (48% of raw data sender/receivers)
- Compute language and attribute similarities, also taking into account user location, for each edge in the graph

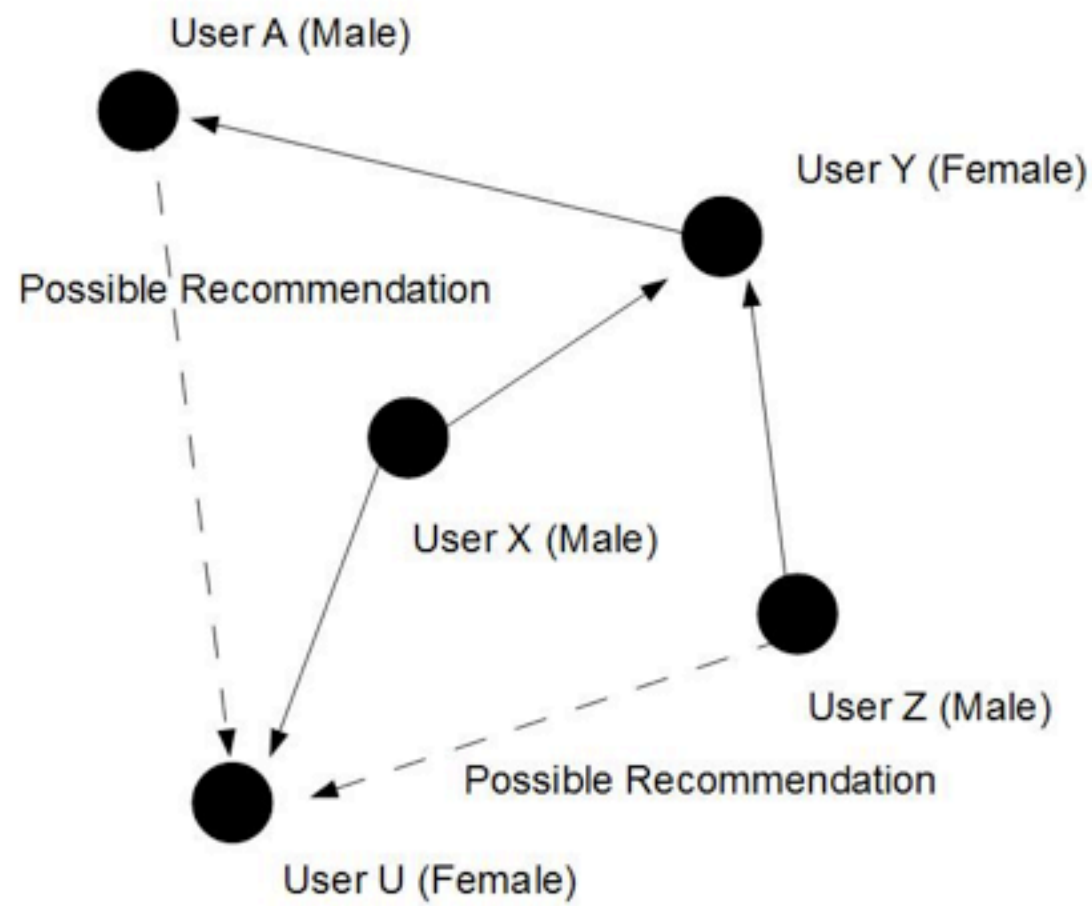
# Message Exchange Modeling

For each user in the communication graph, build a predictive model which can identify successful message exchanges

- Model is based on user's language and attribute similarities with all partners
- Data smoothing is used on independent variables (reduce noise)
- Positive identification based on attribute (content) similarity of at least 80%
- Most important independent variables in the user model are ranked

# Finding Candidates

- Original number of possible (directional) connections population:
  - 532,058 * 532,058 = 283,085,715,364
  - (532,058 * 532,057) / 2 = 141,542,591,653
- Construct FOAF-like graph
  - Population: 218,307 users
  - Directional connections: 11,414,012 (vs 47,657,946,249)






Topicmarks

# Recommendation Methodology

- Each user in the graph of possible recommendations has a classifier.
- A classifier returns 1 if the attributes of another matches the user's.

- So, for each user for each possible candidate to recommen
  - apply user's classifier against candidate's attributes {0, 1}
  - apply candidate's classifier against user attributes {0, 1}

One-way Acceptance => user classifier returns 1 for candidate

Mutual Acceptance => user classifier returns 1 for candidate AND candidate's classifier returns 1 for user



Topicmarks

# Recommendations

| Recommendations | Connections | | | | Users | |
|---|---|---|---|---|---|---|
| | Number | % of FOAF-like Graph | % of Original Data | | Number | % of Original Population |
| **Single Acceptance** | 3,780,544 | 33.12% | 122.20% | (a) | 127,867 | 24.03% |
| **Mutual Acceptance** | 1,621,522 | 14.21% | 58.98% | (b) | 68,973 | 12.96% |
| Population | | 11,413,971 | 1,088,099 | | | 532,058 |
| **Messages Out >= 10** | | | | | | |
| **Single Acceptance** | 859,472 | 36.03% | 1131.94% | | 18,806 | 45.50% |
| **Mutual Acceptance** | 488,048 | 20.46% | 642.77% | | 14,072 | 34.04% |
| Population | | 2,385,260 | 75,929 | | | 41,334 |
| **TextLength Out >= 100** | | | | | | |
| **Single Acceptance** | 941,180 | 33.74% | 1045.58% | | 26,092 | 41.56% |
| **Mutual Acceptance** | 510,504 | 18.30% | 567.13% | | 17,701 | 28.19% |
| Population | | 2,789,338 | 90,015 | | | 62,786 |
| Note: Only California sending recommendations compared to original data. | | | | | | |
| (a) | 1,329,640 | | | | | |
| (b) | 641,778 | | | | | |



Topicmarks

# Conclusion

IM Content analysis has the potential to increase the number of connections between users by more than 50%, impacting 25% of the Tagged community.

# Team

**Peter Berger, CEO**

- Alitora Systems
- WikiLoan
- HotShot Media
- SV SemTech Grp
- UCSF Catalyst T1
- Adams Nye LLP

**Karl Dawson, CTO**

- PhiScape
- DeltaVista
- Rentenanstalt
- iFace
- Credit Suisse
- Amdahl

**Jaromir Dzialo, Engineer**

- PhiScape
- AMS/CGI
- Sabre
- DeltaVista
- Onet.pl
- Microsoft

**Matt Walters, Bus. and Corp. Dev.**

- LifeGivingForce
- IBM
- Room to Read
- AIG
- Wharton
- Princeton

**Pitor Metel, Engineer**

- 20 years of experience
- Distributed computing
- Web applications
- Applications frameworks
- Natural language processing
- Applied Mathematics, Jagiellonian University

**Wojciech Pater, Engineer**

- 10 years experience
- Server-side
- Distributed systems
- Natural language processing fields



Topicmarks

# Contact

Peter Berger, CEO
peter@topicmarks.com

Karl Dawson, CTO
karl@topicmarks.com

Matt Walters, Head of Customer Dev.
matt@topicmarks.com

Topicmarks, Inc.
153 Townsend Street, Suit 900
San Francisco, CA 94107
(415) 310-4406