THE UNIVERSITY of TENNESSEE KNOXVILLE

Using Large-Scale Matrix Factorizations to identify users of Social Networks

Dr. Michael W. Berry and Denise Koessler

In celebration of Robert J. Plemmons 75th Birthday

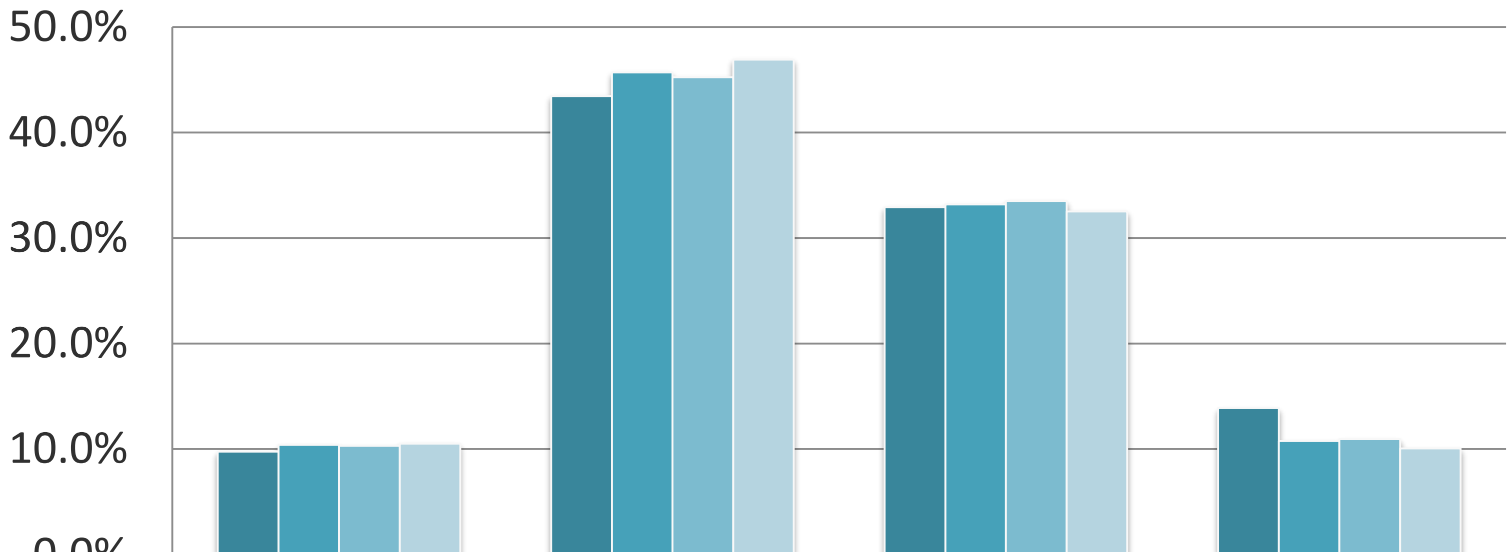
The Chinese University of Hong Kong

November 17, 2013

BIG ORANGE BIG IDEAS

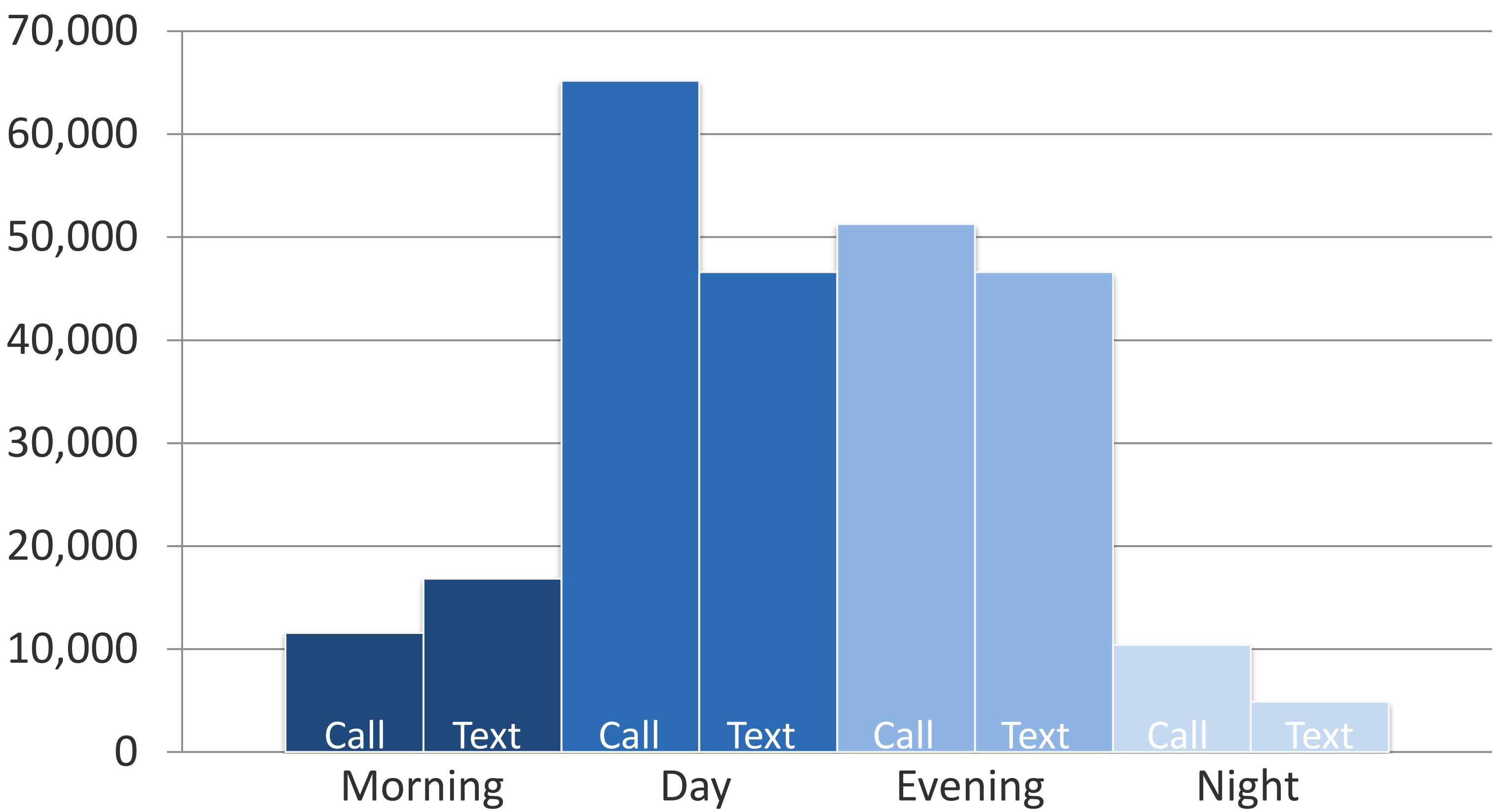
CISML
CENTER FOR INTELLIGENT SYSTEMS AND MACHINE LEARNING
THE UNIVERSITY of TENNESSEE COLLEGE of ENGINEERING

Percent of total calling behavior observed in four different cities during time *t*



	Morning Calls	Day Calls	Evening Calls	Night Calls
City A	9.8%	43.5%	32.9%	13.9%
City B	10.4%	45.7%	33.2%	10.8%
City C	10.3%	45.2%	33.5%	10.9%
City D	10.5%	46.9%	32.5%	10.1%

Number of users who spend more than 25% of their total activity during time *t*



THE UNIVERSITY of TENNESSEE KNOXVILLE

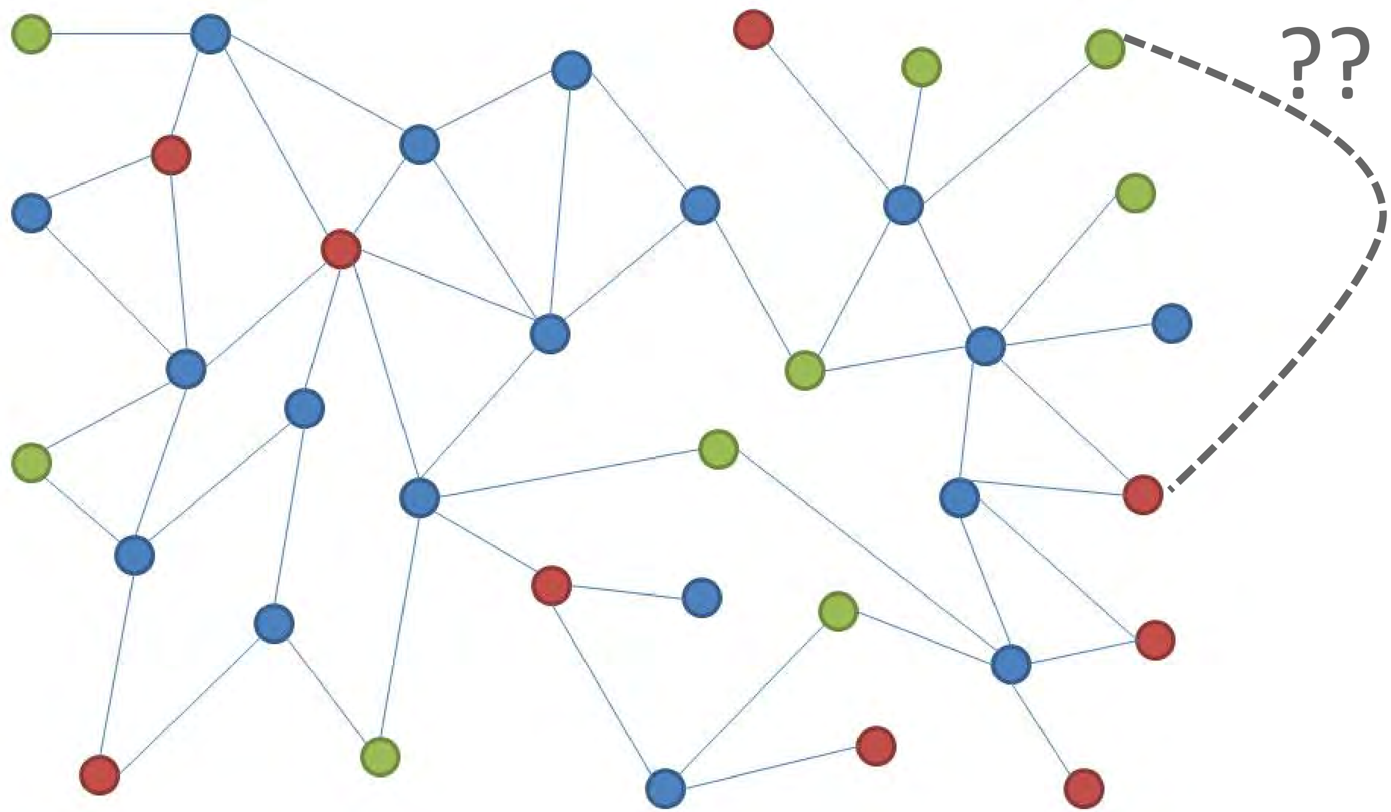
BIG ORANGE BIG IDEAS

Is a mobile customer's mobile behavior unique? Yes

Yves et. al, Unique In the Crowd, March 2013, *Nature*

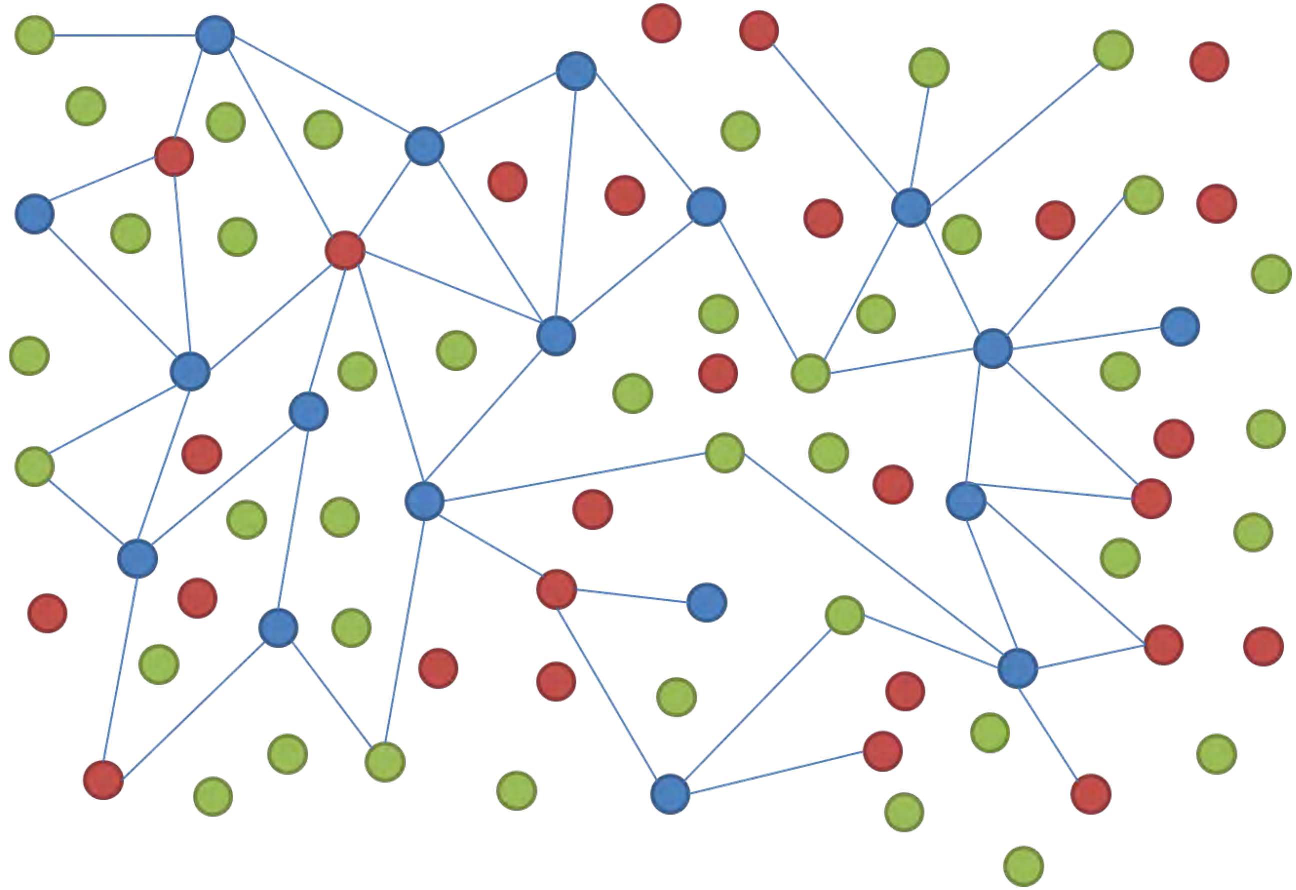
Do we need *physical* location?

Why is this difficult?



Why is this difficult?

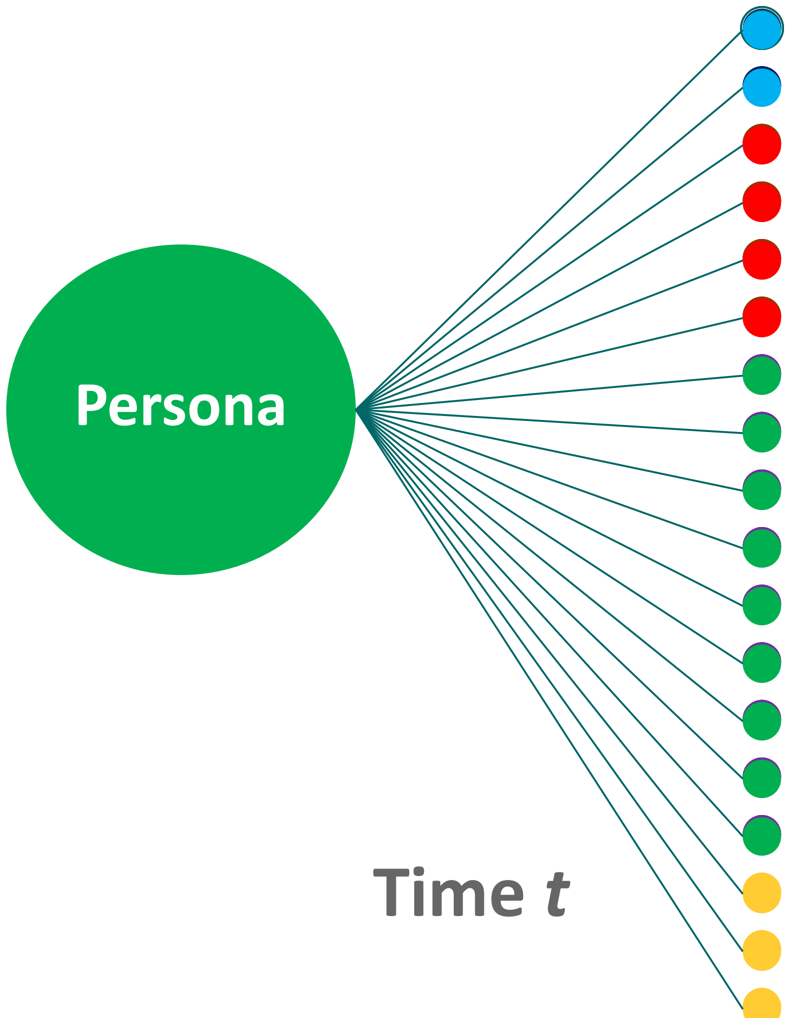
The actual world...



Research Goal:

Given a social network, can we detect key components of user data that uniquely identifies individuals throughout time?

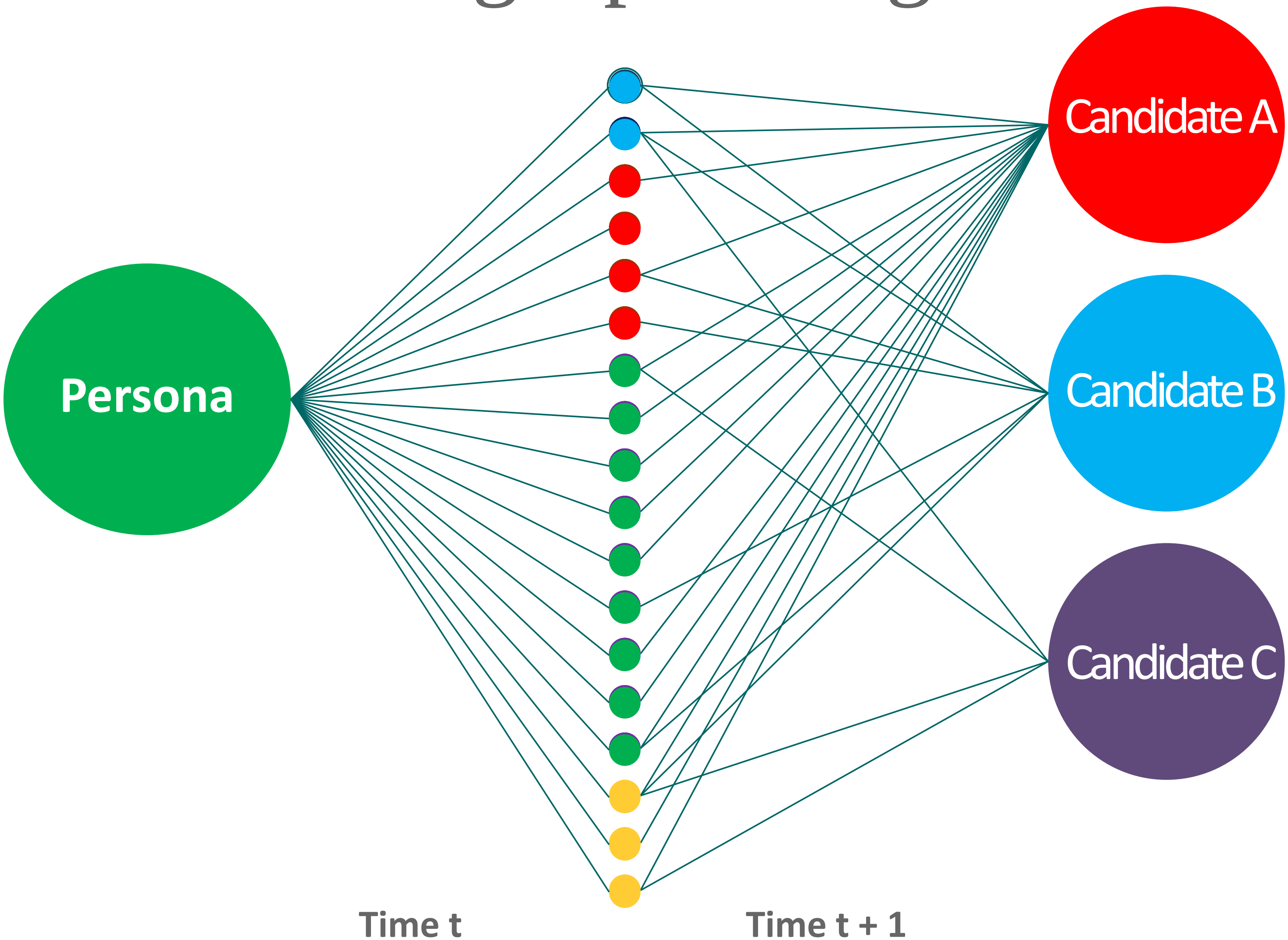
Preliminary Approaches: Social Fingerprinting



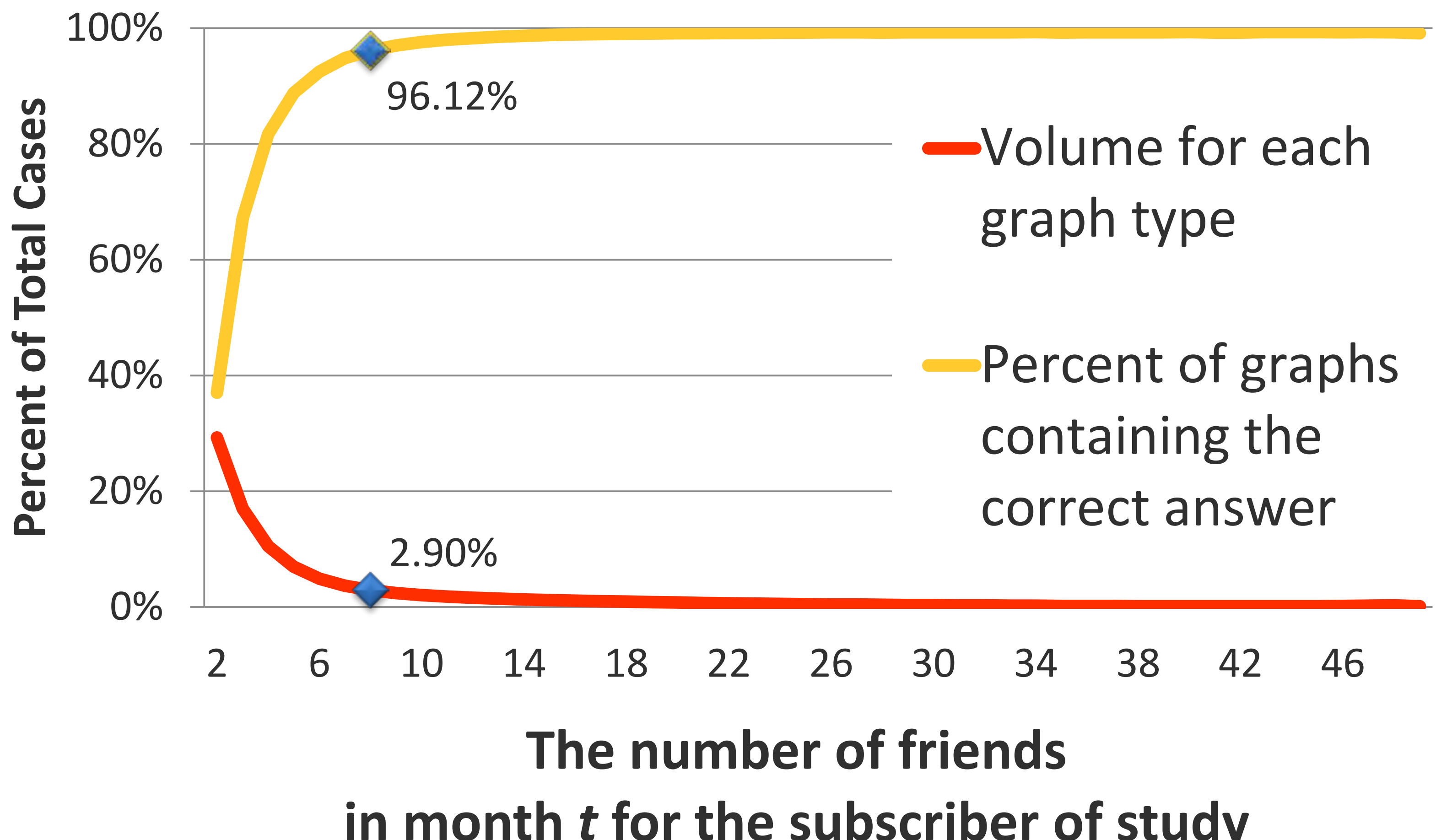
Goal:

Accurately identify social network users based on features of a dynamic, labeled graph

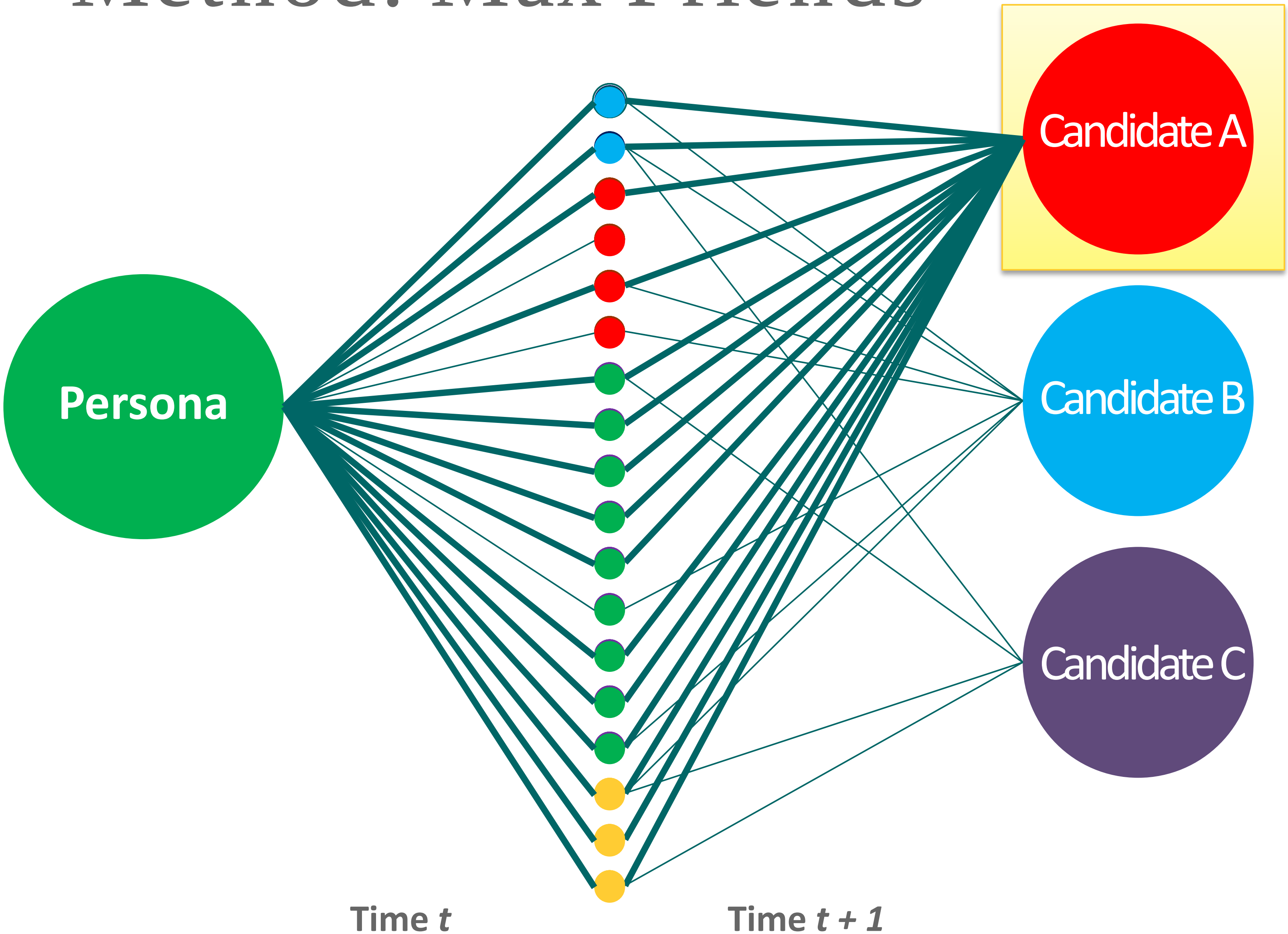
Social Fingerprinting



Statistics for second neighbor graphs: created from one month of history

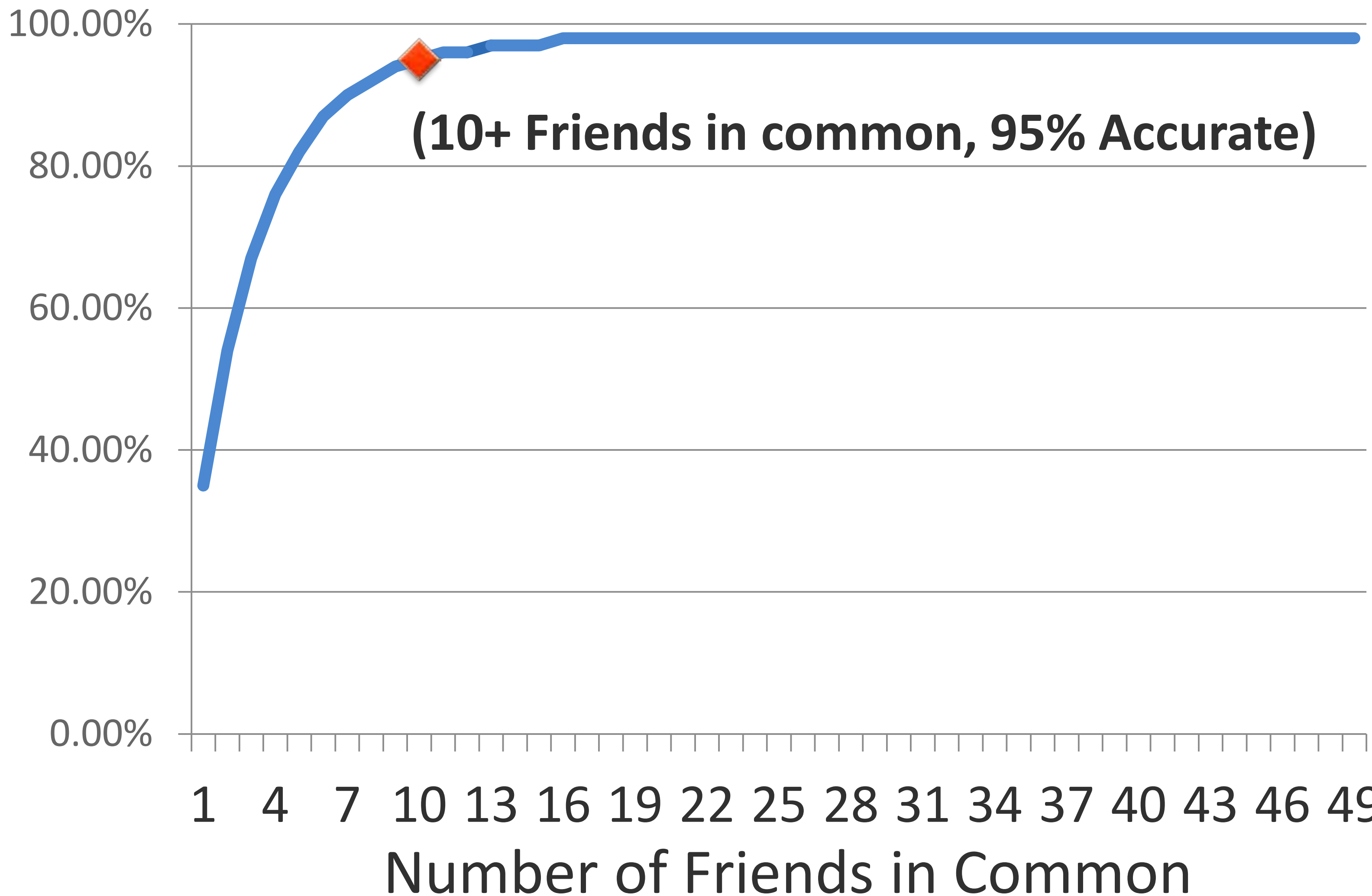


Method: Max Friends



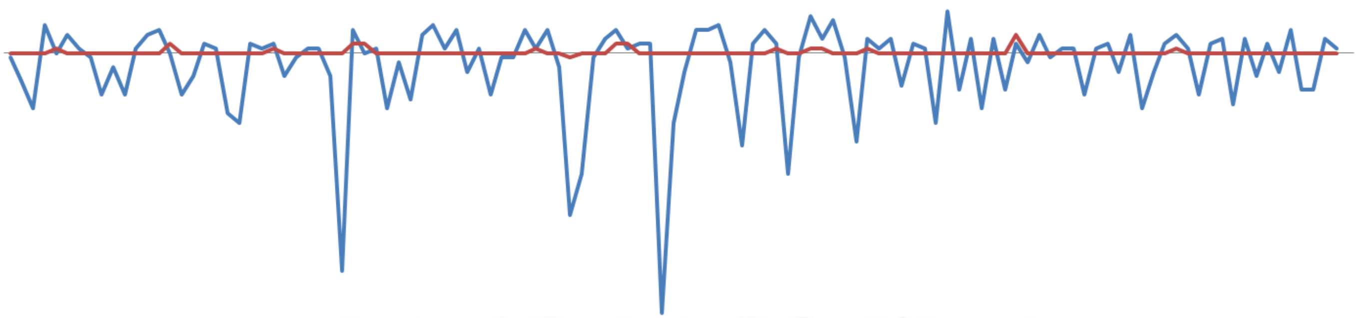
Accuracy Max Friends

One Month of History

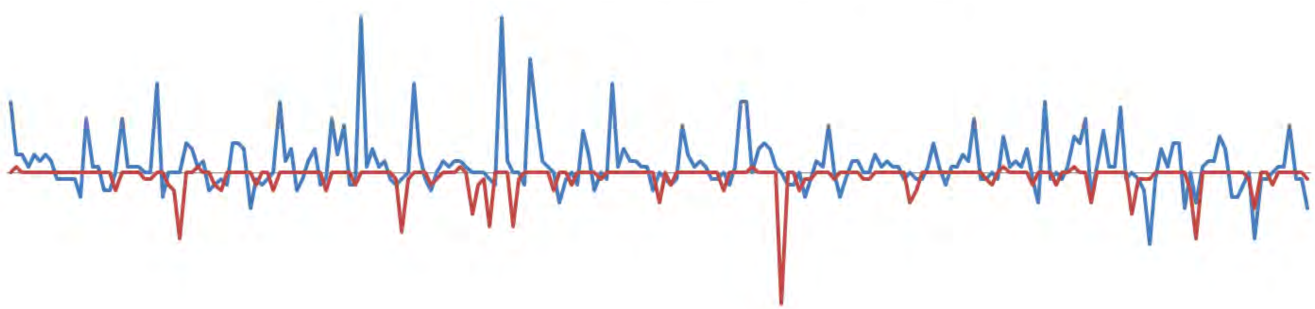


Need: identification of features

Social Network User A



Social Network User B



Semidiscrete Decomposition (SDD) [Kolda and O'Leary 1998]

$$A_{nxn} \approx U_{nxk} \cdot \Sigma_{kxk} \cdot V_{nxk}$$

$$U_{i,j} \in \{-1,0,1\}$$

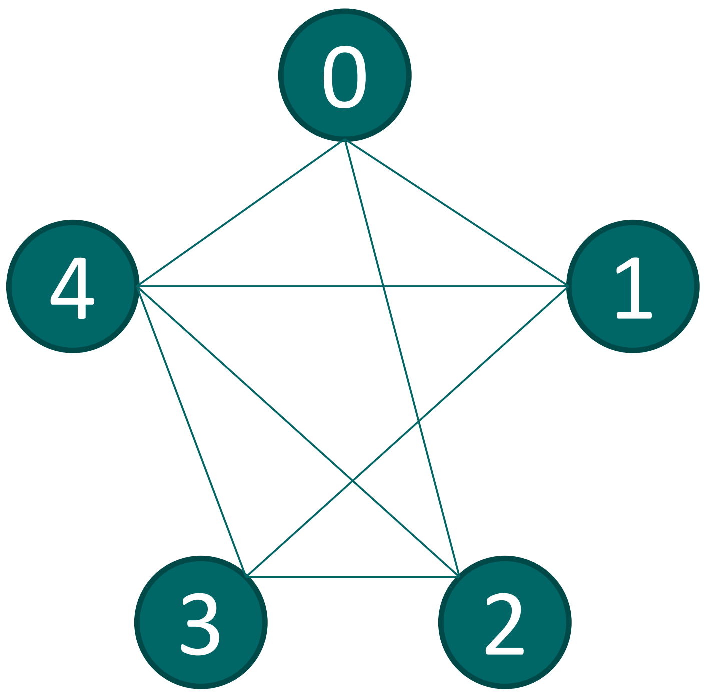
$$\Sigma_{i,i} = \sigma_i$$

$$V_{i,j} \in \{-1,0,1\}$$

SDD Procedure:

1. Construct matrix A and query vector(s)
2. Semidiscrete Decomposition of matrix A to yield rank-*k* approximation
3. Compute new query vector
4. Rank the personas wrt cosine similarity
5. Evaluate

Construction:

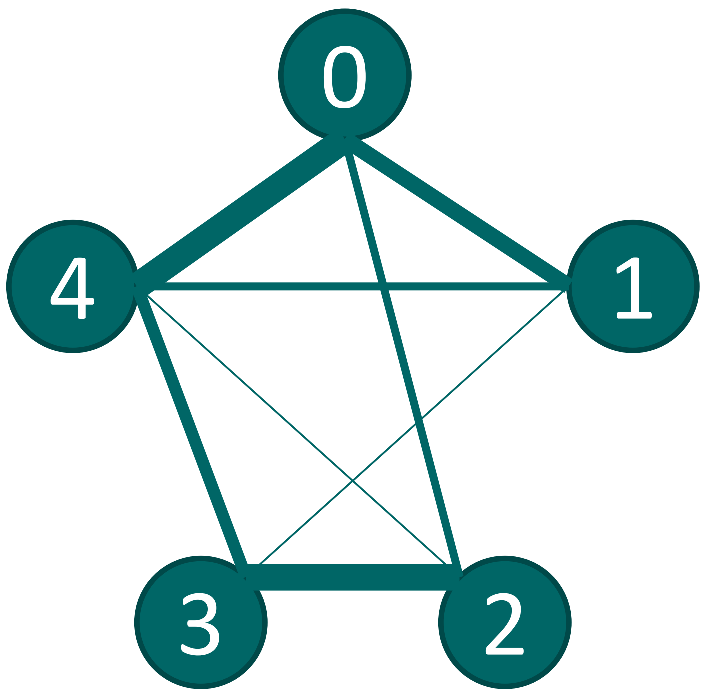


Time *t*

$$
\begin{array}{c}
 \\ p_0 \\ p_1 \\ p_2 \\ p_3 \\ p_4
\end{array}
\begin{array}{c}
\begin{array}{ccccc} p_0 & p_1 & p_2 & p_3 & p_4 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \end{array} \\
\begin{bmatrix}
0 & 1 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 1 \\
1 & 0 & 0 & 1 & 1 \\
0 & 1 & 1 & 0 & 1 \\
1 & 1 & 1 & 1 & 0
\end{bmatrix}
\end{array}
$$

Binary: *A[i,j]* → the presence of the relationship

Construction:

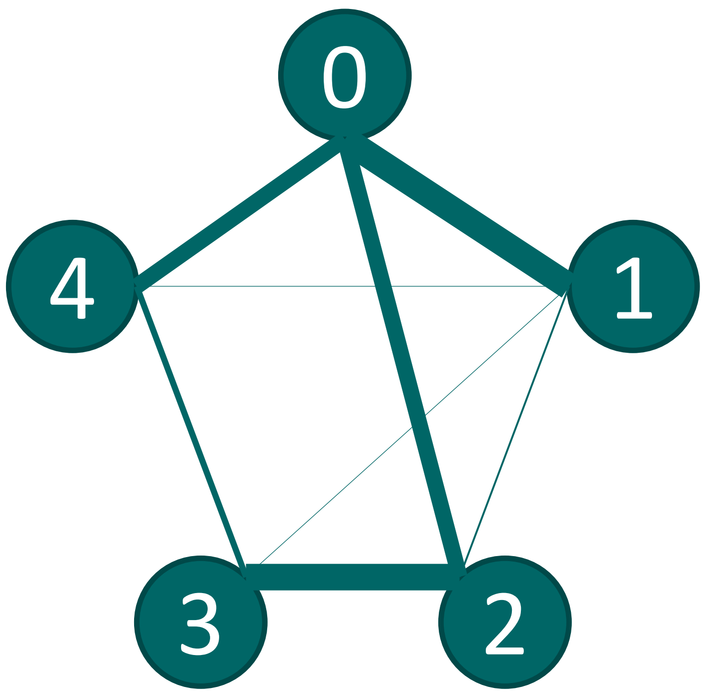


Time *t*

$$
\begin{array}{c|ccccc}
 & p_0 & p_1 & p_2 & p_3 & p_4 \\
 & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\hline
p_0 & 0 & 5 & 2 & 0 & 9 \\
p_1 & 5 & 0 & 0 & 1 & 2 \\
p_2 & 2 & 0 & 0 & 9 & 1 \\
p_3 & 0 & 1 & 9 & 0 & 4 \\
p_4 & 9 & 2 & 1 & 4 & 0
\end{array}
$$

Term Frequency: *A[i,j]* → the *strength* of the relationship

Construction: Query Vectors



Time $t + 1$

$$
\begin{array}{ccccc}
q_0 & q_1 & q_2 & q_3 & q_4 \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\begin{bmatrix}0\\9\\4\\0\\3\end{bmatrix} &
\begin{bmatrix}9\\0\\1\\1\\1\end{bmatrix} &
\begin{bmatrix}1\\3\\0\\9\\0\end{bmatrix} &
\begin{bmatrix}0\\1\\9\\0\\2\end{bmatrix} &
\begin{bmatrix}3\\1\\0\\2\\0\end{bmatrix}
\end{array}
$$

SDD of A: k = 3

$$A_{nxn} \approx U_{nxk} * \Sigma_{kxk} * V_{nxk}$$

$$A_{5x5} \approx \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 7 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 6.5 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

SDD of A: $k = 3$

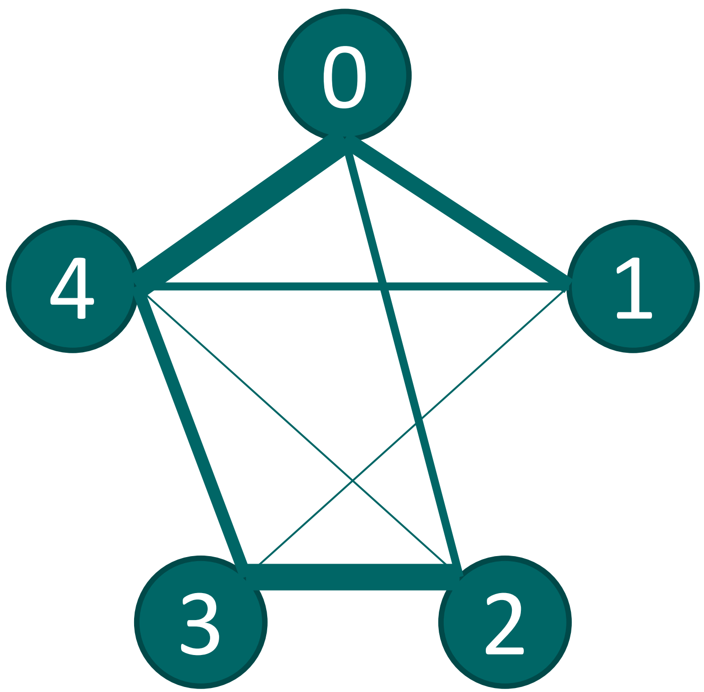
```
Iteration Residual Improvement           Inner  Total
  Number   Squared    (beta)              Its   InnerIts
--------- -------- -----------           ----- --------
    1          1    3.28e+02 9.80000e+01   2        2
    2          3    1.78e+02 1.50000e+02   3        5
    3          4    9.35e+01 8.45000e+01   2        7

    -- SDD information --
final residual norm          : 9.6695e+00
final relative residual norm: 0.468
total outer iterations       : 3
average inner iterations     : 2.333
average init iterations      : 1.333
```

Query Vector Reduction

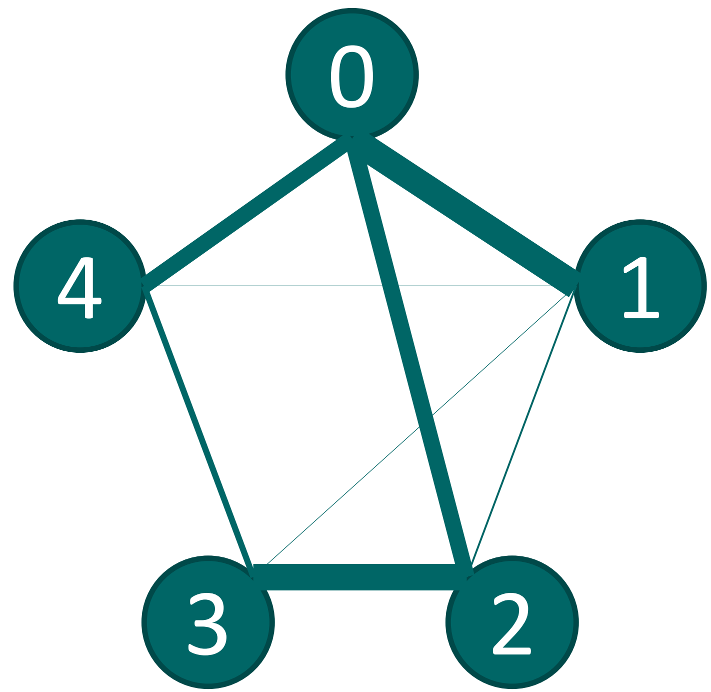
$$q = q^T U_k \Sigma^{-1}$$

$$q_0 \rightarrow [1.71429 \quad 0 \quad 1.07692]$$
$$q_1 \rightarrow [0.14857 \quad 2 \quad 0.30769]$$
$$q_2 \rightarrow [0.42857 \quad 2 \quad 0]$$
$$q_3 \rightarrow [0.42857 \quad 0 \quad 1.69231]$$
$$q_4 \rightarrow [0.14857 \quad 1 \quad 0]$$

Similarity between these graphs:



Time *t*



Time *t + 1*

Cosine Similarity: $q^{t+1}[j]*V^{(t)}[i]$

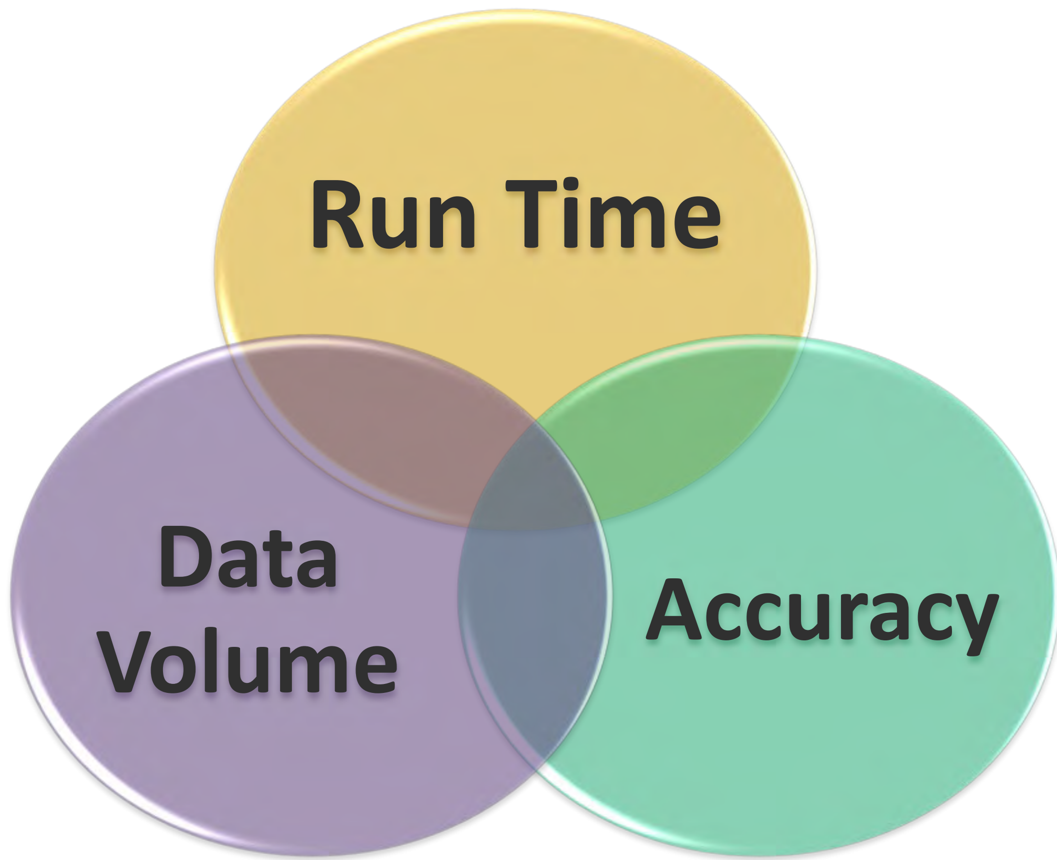
	V[0]	V[1]	V[2]	V[3]	V[4]
q[0]	**0.8467**	0	0	0.5319	0
q[1]	0.0704	**0.9859**	0.9859	0.1516	0.9859
q[2]	0.2095	**0.9778**	**0.9778**	0	**0.9778**
q[3]	0.2454	0	0	**0.9693**	0
q[4]	0.1414	**0.9899**	**0.9899**	0	**0.9899**

Future work using SDD:

1. An optimal parameter *k*?
2. Additional similarity measures
3. How often is a persona ranked in the top 1%?
4. When this approach is incorrect, what does the distribution of the correct identity look like?
5. Is there a threshold for inconclusively?
6. Find a confidence factor → is there a large separation in scores?

Conclusions

We have a triad of issues:



Conclusions from a **Big Data Perspective**:

At this point, we are either:

- Accurate on a small portion of the data on any window of time.
- Accurate on all of the data given infinite amount of storage space

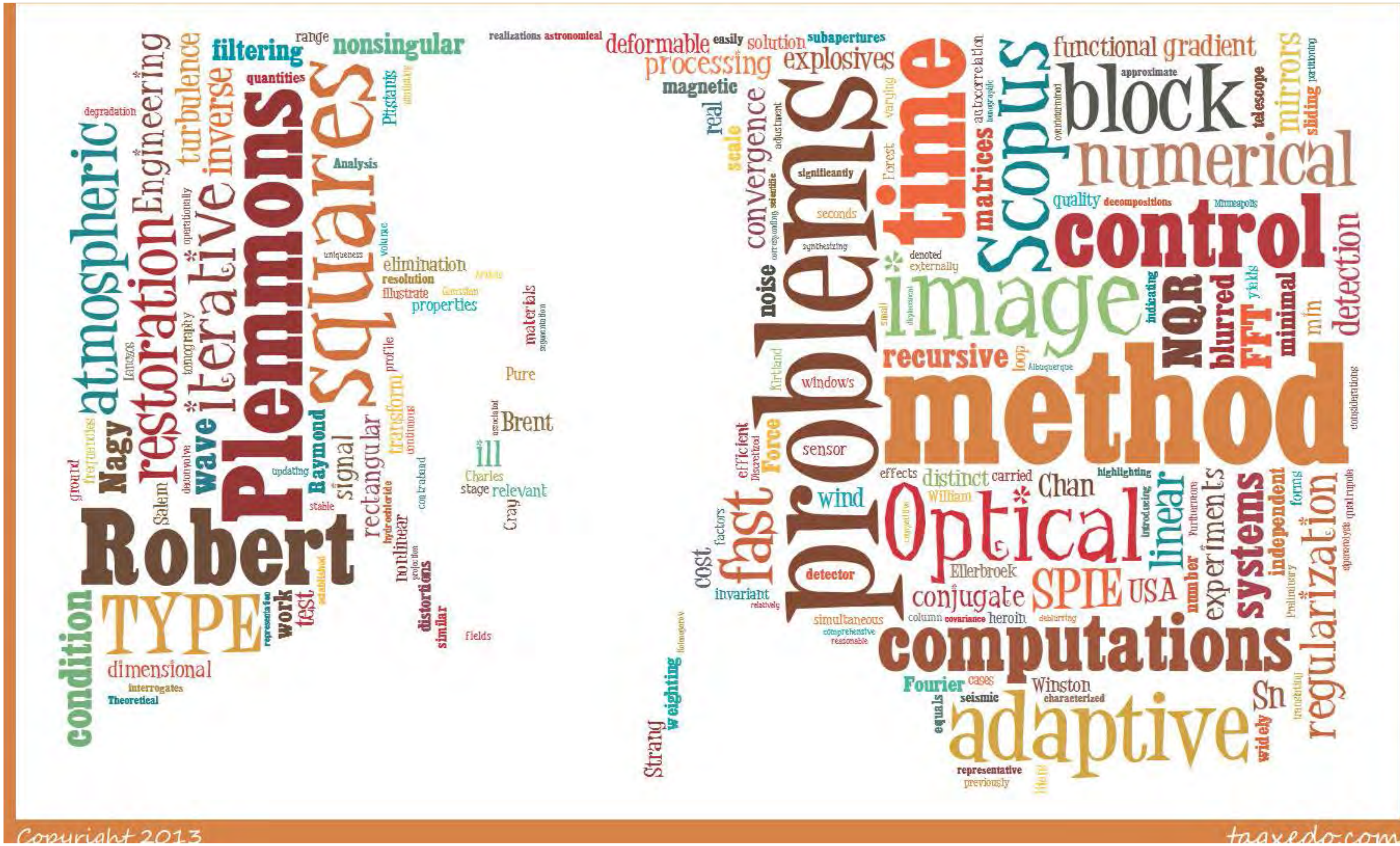
… or …

- Able to classify volumes of social inferences in real time with low confidence.

References

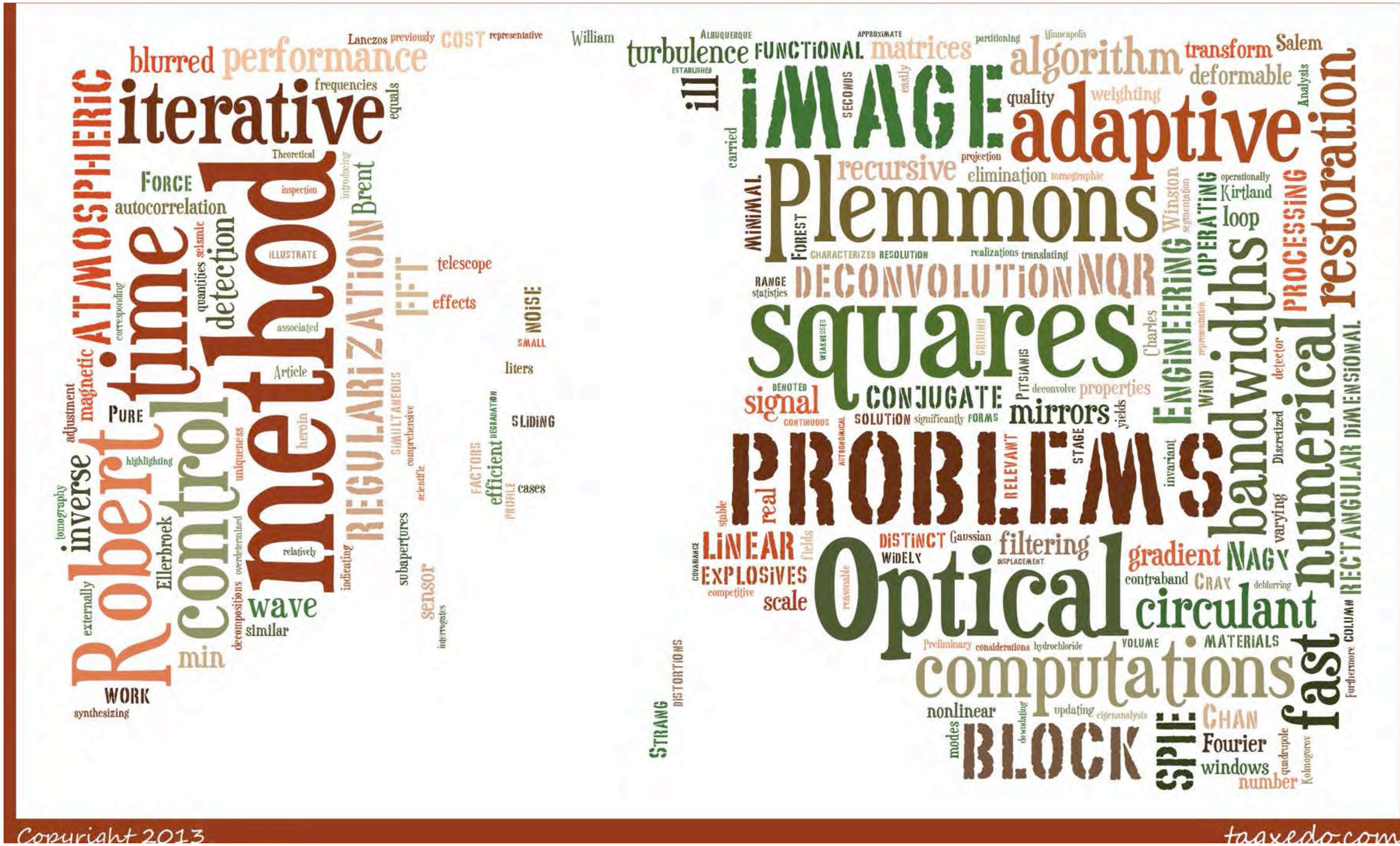
- R. Becker, C. Volinsky, and A. Wilks. 2010. Fraud Detection in Telecommunications History and Lessons Learned. In *Technometrics*. Vol. 52, No 1.
- C. Cortes, D. Pregibon, and C. Volinsky. 2001. Communities of Interest. In*Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis* (IDA '01). Springer-Verlag, London, UK, UK, 105-114.
- S. Keshav. 2005. Why cell phones will dominate the future internet. SIGCOMM Comput. Commun. Rev. 35, 2 (April 2005), 83-86. DOI=10.1145/1064413.1064425 http://doi.acm.org/10.1145/1064413.1064425.
- A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi. 2006. On the structural properties of massive telecom call graphs: findings and implications. In *Proceedings of the 15th ACM international conference on Information and knowledge management* (CIKM '06). ACM, New York, NY, USA, 435-444. DOI=10.1145/1183614.1183678 http://doi.acm.org/10.1145/1183614.1183678
- J. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J.Kertesz, and A. Barabasi. 2007. Structure and tie strengths in mobile communication networks. In *PNAS*. Vol 104. No. 18. 7332 – 7336.
- X. Ying and X. Wu. 2009. On Randomness Measures for Social Networks. In *SIAM International Conference on Data Mining*. 709 – 720.

BIG ORANGE BIG IDEAS



BIG ORANGE BIG IDEAS

BIG ORANGE BIG IDEAS




METHOD
Optical
Proceedings
SPIE
squares
time
NUMERICAL
TYPE
inverse
fast
convergence
RECURSIVE
transform
Strang
detection
noise
matrices
MINIMAL
number
Brent
Winston
autocorrelation
algorithm
matrix
problems
image
control
circulant
block
Robert
RECTANGULAR
MIRRORS
preconditioners
computations
PLEMMONS
ADAPTIVE
ITERATIVE
BANDWIDTHS
Toeplitz
atmospheric
Nagy
NQR
FFT
wave
linear
processing
TURBULENCE
experiments
nonsingular
signal
filtering
Chan
EXPLOSIVES
GRADIENT
condition
blurred
Cray
cost
Force
Raymond
elimination
INDEPENDENT

Extra slides follow..

Ranking Alternatives:

Structure A and q:

1) Persona x Persona
2) Persona x Time
3) Persona x Persona x Time

SDD

Select Ranking Function:

1) Cosine
2) Euclidean
3) Jaccard
4) Pearson

Evaluate Performance