THE UNIVERSITY of TENNESSEE KNOXVILLE

Social Fingerprinting: Identifying Users of Social Networks by their Data Footprint

The University of Tennessee
Electrical Engineering and Computer Science
Dissertation Defense

Denise Koessler Gosnell

THE UNIVERSITY of TENNESSEE KNOXVILLE

Is a social network user's behavior unique?

THE UNIVERSITY of TENNESSEE KNOXVILLE

Has social data made anonymity impossible?

87%

U.S. adults' location is known via their mobile phone

MIT T.R.

U.S. young adults carry their phone wherever they go

AT&T

96%

3

THE UNIVERSITY of TENNESSEE KNOXVILLE

Outline:

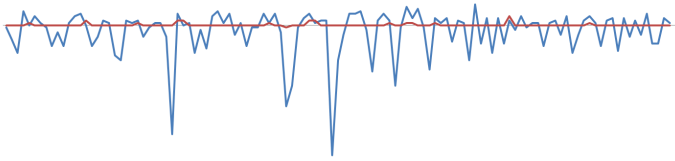
1. Define and Motivate
2. Model and Software
3. Individual Printing Algorithm
4. Community Printing Algorithm
5. Conclusions

4

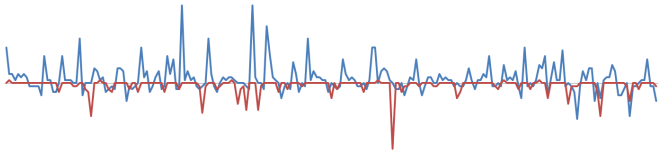
THE UNIVERSITY of TENNESSEE KNOXVILLE

A Social Fingerprint

Social Network User A



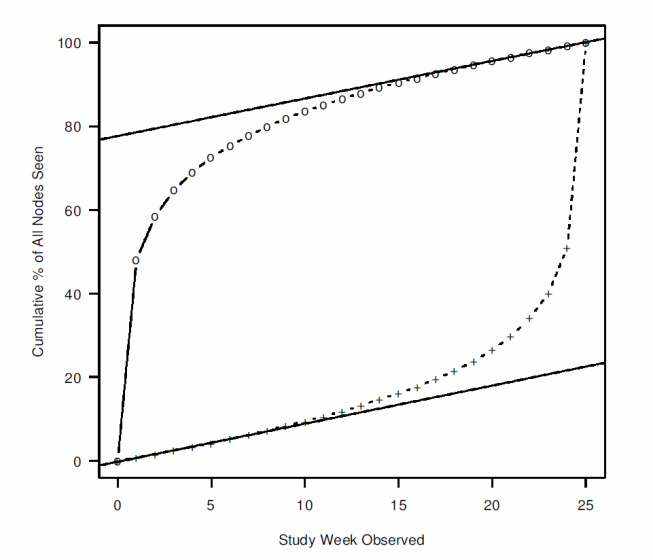
Social Network User B



5

THE UNIVERSITY of TENNESSEE KNOXVILLE

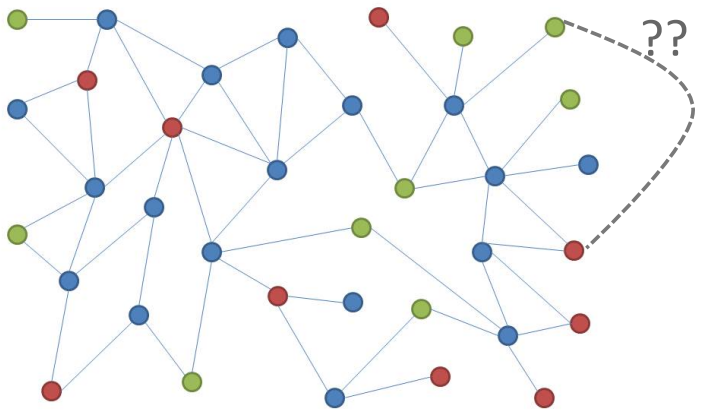
Why is this hard?



6

THE UNIVERSITY of TENNESSEE KNOXVILLE

Why is this hard?



7

THE UNIVERSITY of TENNESSEE KNOXVILLE

Previous Work

ACADEMIA

- Bounds of Human Privacy, MIT
- SNAP Project, Stanford
- Graph Modeling Libraries
- Simulated topologies and statistics

INDUSTRY

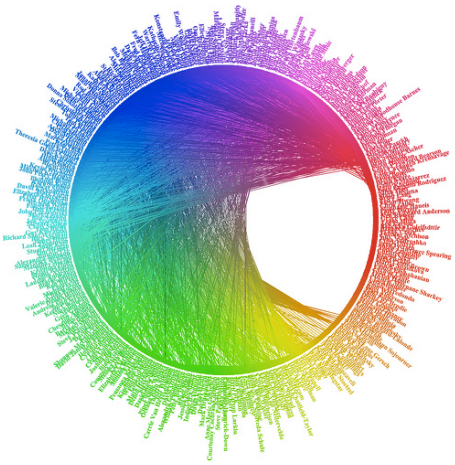
- IBM Watson SMSim
- Topology Statistics
- Private Companies
- …?

8

THE UNIVERSITY of TENNESSEE KNOXVILLE

The Model

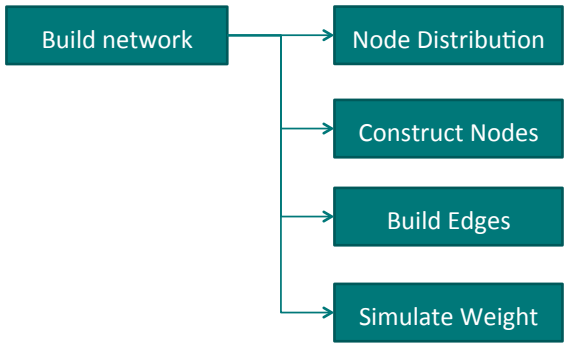
- Connectivity?
- Time?
- Multiple Edges?
- Empirical Data vs. Assumptions?



9

THE UNIVERSITY of TENNESSEE KNOXVILLE

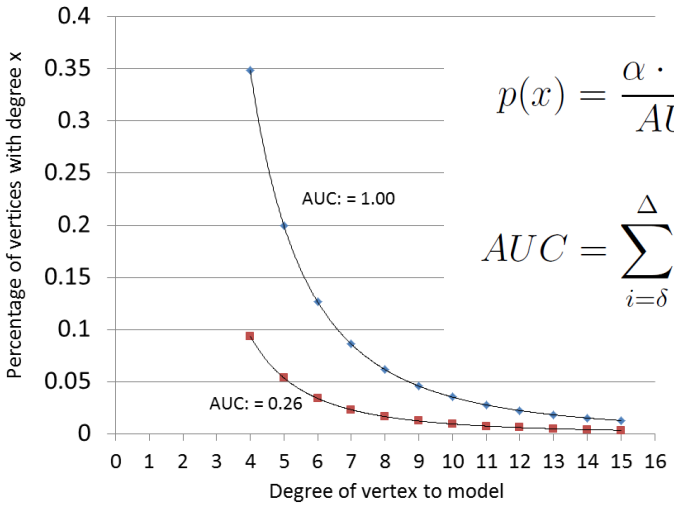
The Model



10

THE UNIVERSITY of TENNESSEE KNOXVILLE

Scale-free Distribution



$$p(x) = \frac{\alpha \cdot x^{-\beta}}{AUC}$$

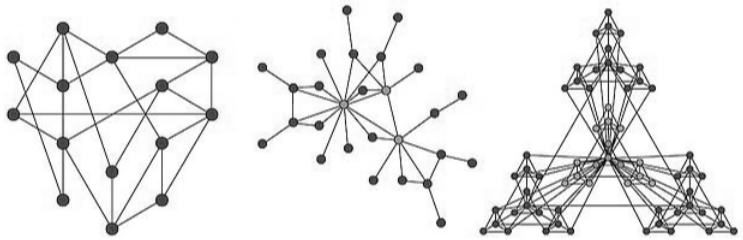
$$AUC = \sum_{i=\delta}^{\Delta} \alpha \cdot i^{-\beta}$$

11

THE UNIVERSITY of TENNESSEE KNOXVILLE

The Base Model

Scale-Free Graph



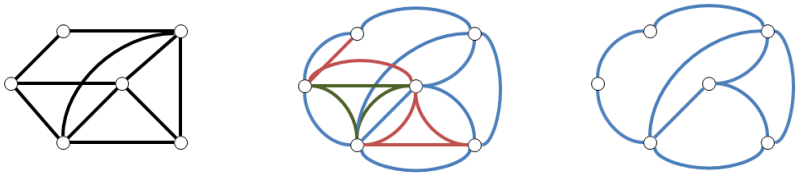
Random Graph

Hierarchical Graph

12

THE UNIVERSITY of TENNESSEE KNOXVILLE

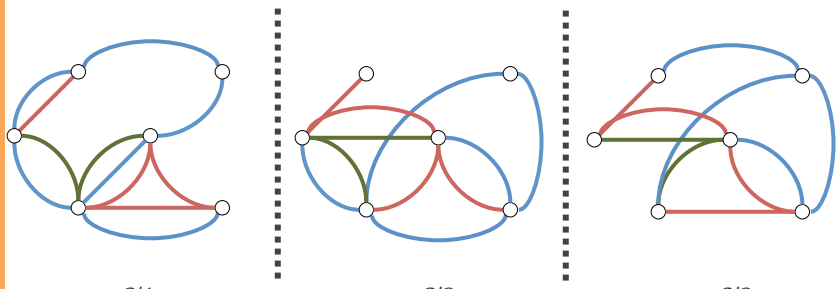
Multiple Edges



13

THE UNIVERSITY of TENNESSEE KNOXVILLE

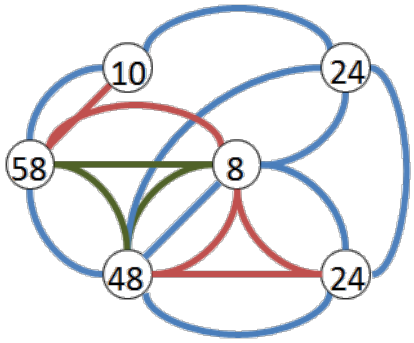
Dynamic Edges



14

THE UNIVERSITY of TENNESSEE KNOXVILLE

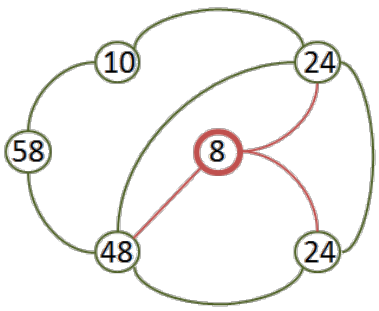
Simulating Edge Weights

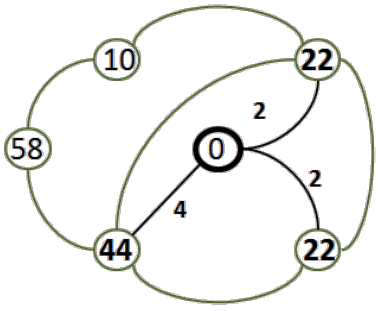


15

THE UNIVERSITY of TENNESSEE KNOXVILLE

Maximal Diffusion Algorithm

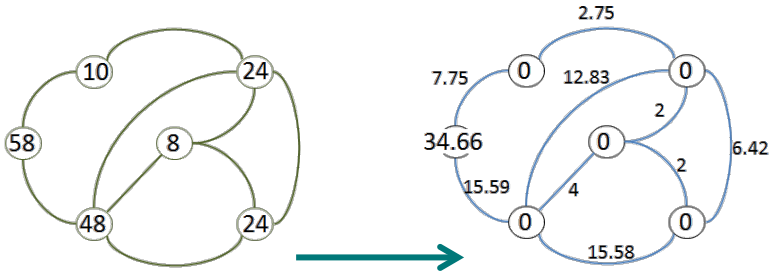




16

THE UNIVERSITY of TENNESSEE KNOXVILLE

Maximal Diffusion Algorithm



Capacity constraints: $\sum_{v_j \in N(v_i)} w(e(i,j)) \leq c_{i,b}$

Flow conservation: $r_{i,b} + \sum_{v_j \in N(v_i)} w(e(i,j)) = c_{i,b}$

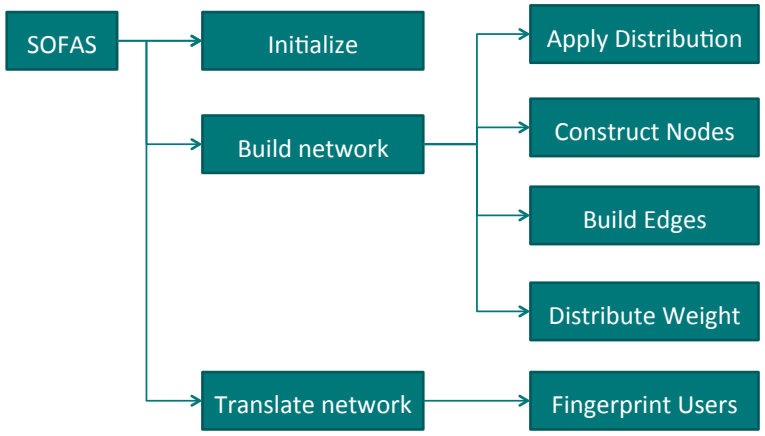
THE UNIVERSITY of TENNESSEE KNOXVILLE

Maximal Diffusion Algorithm

Result: A fractional estimation of an optimally weighted graph via the diffusion of node capacity.
Input: time step t and edge type b;
Sort all nodes such that $c^t_{i,b} \leq c^t_{i+1,b}$;
for *each* $v_i \in V(G)_{sorted}$ **do**
  Initialize: $r^t_{i,b} = c^t_{i,b}$;
  Determine total community capacity C of v_i;
    $C = \sum_{v_j \in N(v_i)} c^t_{j,b}$;
  where $N(v_i) = \{v_j \in V(G^t) | e(i,j)^t_b \in E(G^t)\}$;
  for *each* $v_j \in N(v_i)$ **do**
    $w = c^t_{i,b} \cdot \frac{c^t_{j,b}}{C}$;
    $e(i,j)^t_b = w$;
    $c^t_{j,b} = c^t_{j,b} - w$;
    $r^t_{i,b} = r^t_{i,b} - w$;
  end
  Sort all nodes starting from v_i such that $c^t_{i,b} \leq c^t_{i+1,b}$;
end

THE UNIVERSITY of TENNESSEE KNOXVILLE

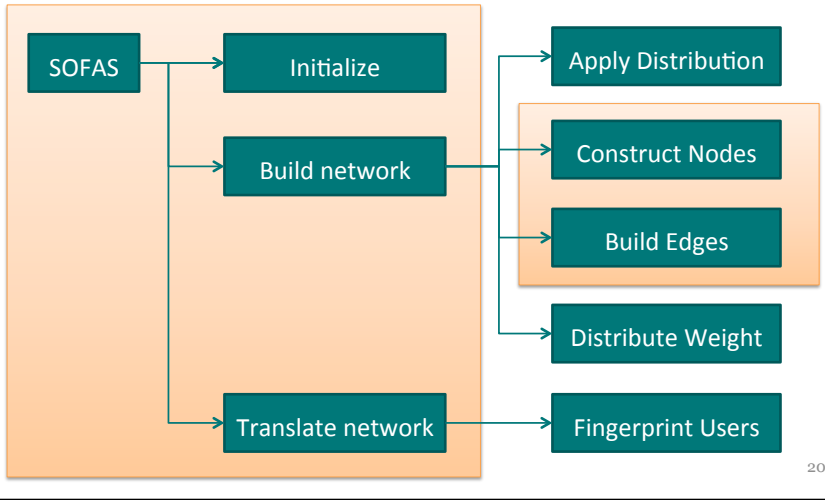
SOcial Fingerprint Analysis Software (SOFAS)



19

THE UNIVERSITY of TENNESSEE KNOXVILLE

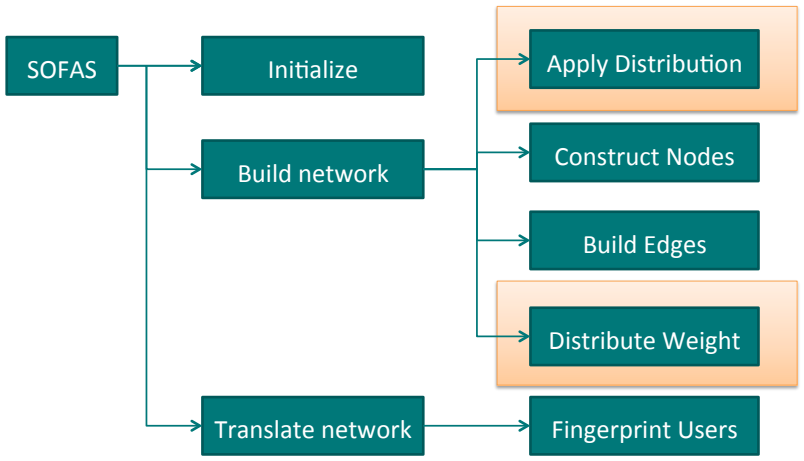
SOFAS: Phases 1 & 2



20

THE UNIVERSITY of TENNESSEE KNOXVILLE

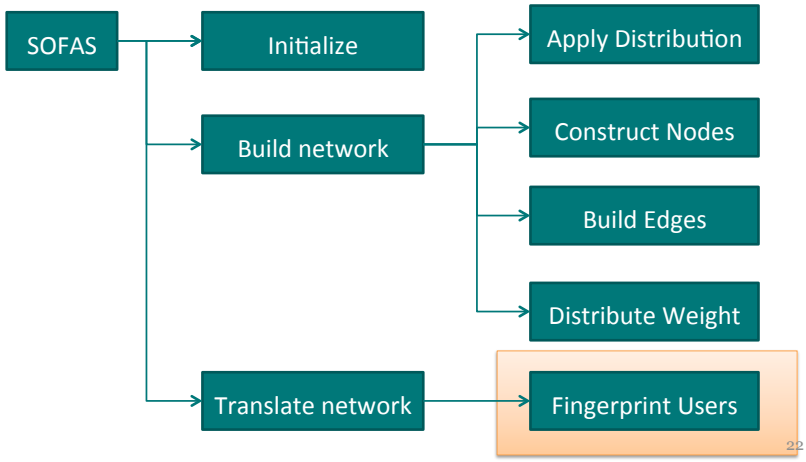
SOFAS: Phases 3 and 4



21

THE UNIVERSITY of TENNESSEE KNOXVILLE

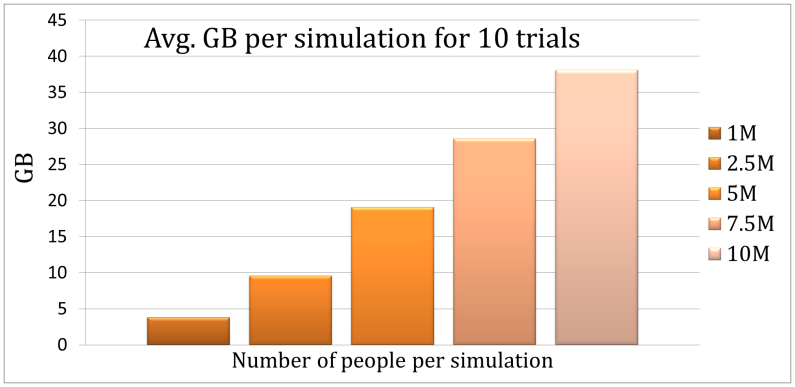
SOFAS: Phase 5



22

THE UNIVERSITY of TENNESSEE KNOXVILLE

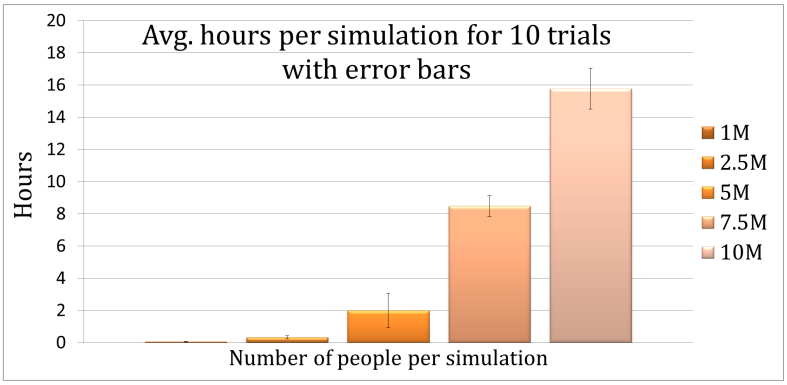
SOFAS: Memory Performance



23

THE UNIVERSITY of TENNESSEE KNOXVILLE

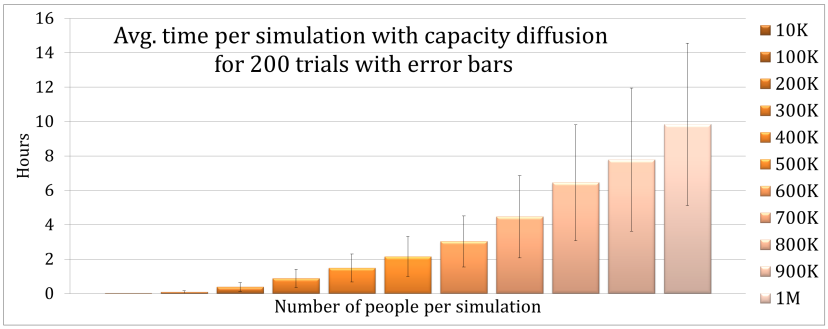
SOFAS: Time



24

THE UNIVERSITY of TENNESSEE KNOXVILLE

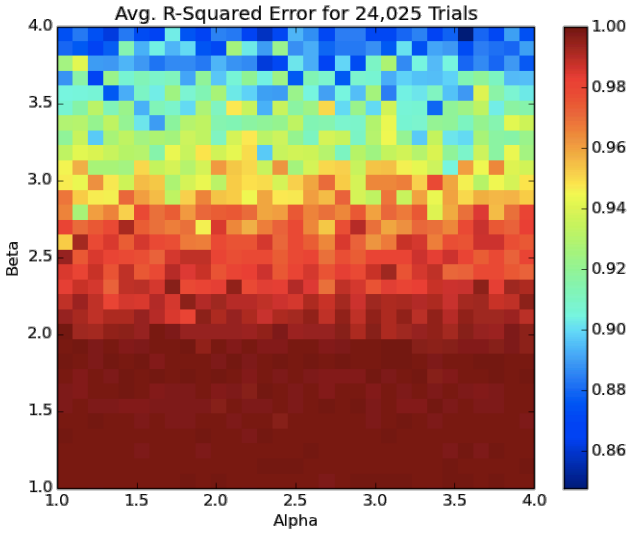
SOFAS: Diffusion Time



25

THE UNIVERSITY of TENNESSEE KNOXVILLE

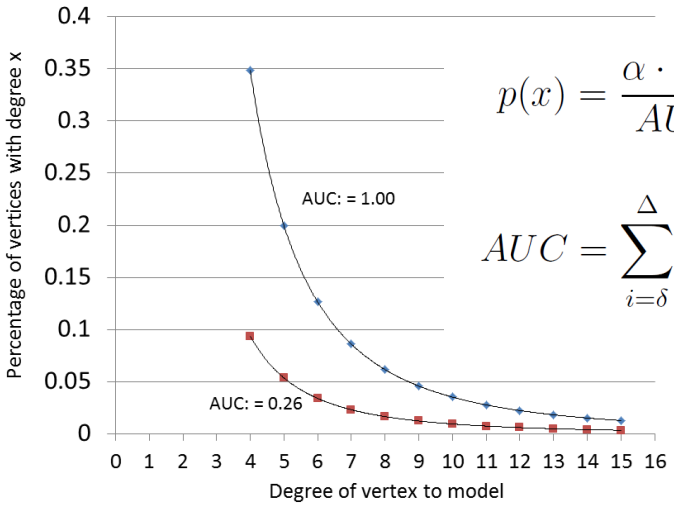
SOFAS: Distribution Error



26

THE UNIVERSITY of TENNESSEE KNOXVILLE

Scale-free Distribution



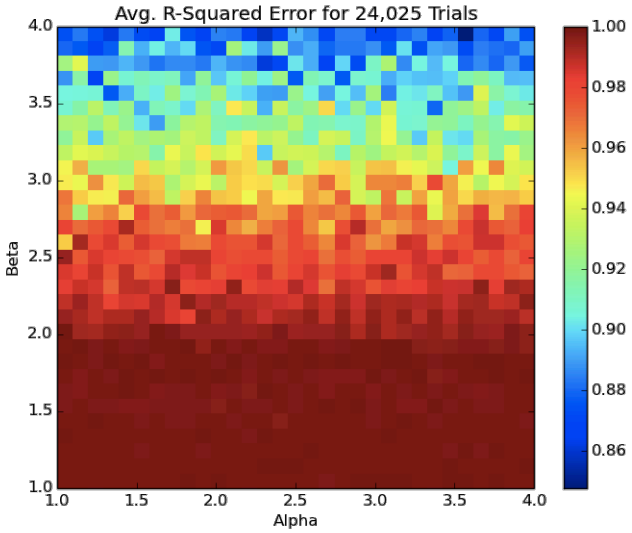
$$p(x) = \frac{\alpha \cdot x^{-\beta}}{AUC}$$

$$AUC = \sum_{i=\delta}^{\Delta} \alpha \cdot i^{-\beta}$$

27

THE UNIVERSITY of TENNESSEE KNOXVILLE

SOFAS: Distribution Error



28

THE UNIVERSITY of TENNESSEE KNOXVILLE

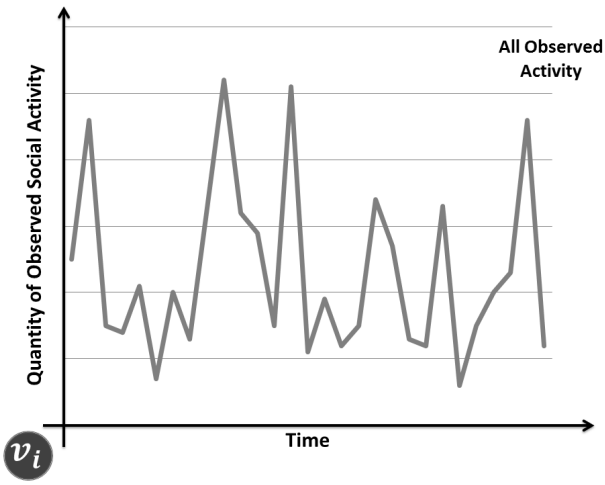
Social Fingerprinting:

Identifying the Individual

29

THE UNIVERSITY of TENNESSEE KNOXVILLE

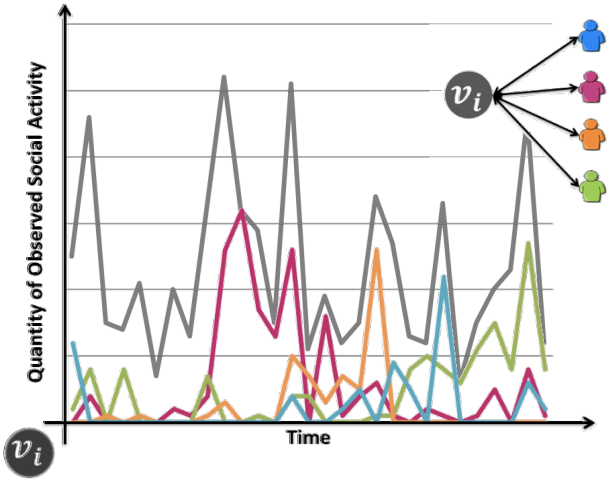
The Theory



30

THE UNIVERSITY of TENNESSEE KNOXVILLE

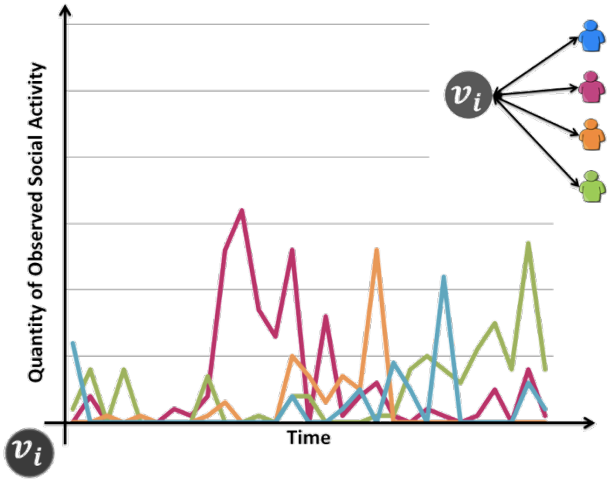
The Theory



31

THE UNIVERSITY of TENNESSEE KNOXVILLE

The Theory



32

THE UNIVERSITY of TENNESSEE KNOXVILLE

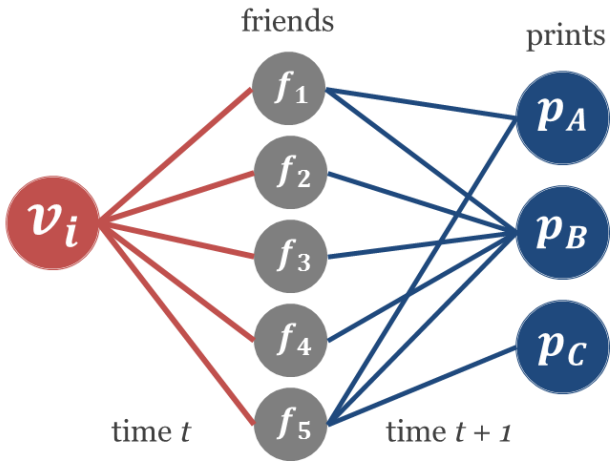
Social Fingerprint Procedure:

1. For each person, construct a neighborhood graph
2. Given the neighborhood graph, determine candidate prints
3. Rank the prints
4. Evaluate

33

THE UNIVERSITY of TENNESSEE KNOXVILLE

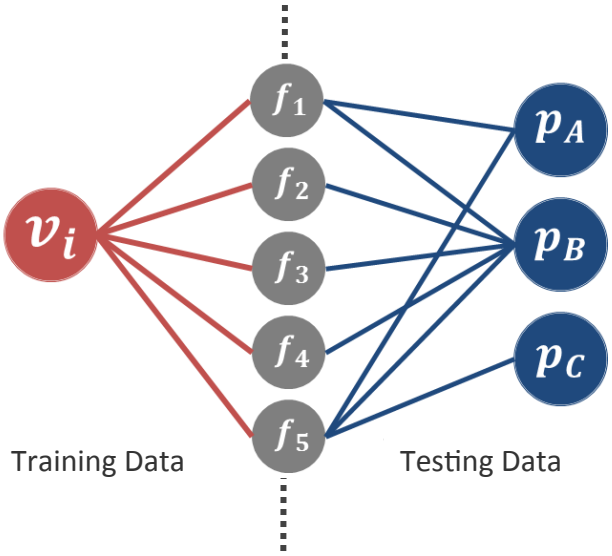
Print Construction



34

THE UNIVERSITY of TENNESSEE KNOXVILLE

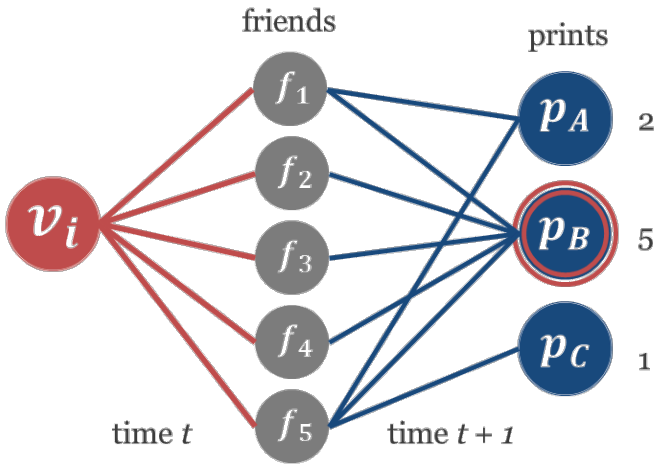
Print Construction



35

THE UNIVERSITY of TENNESSEE KNOXVILLE

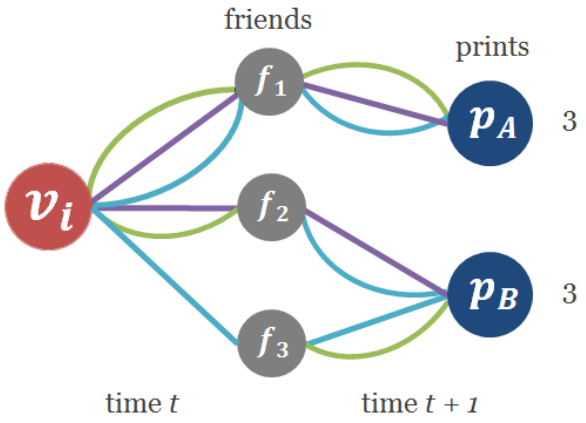
Intersection Score



36

THE UNIVERSITY of TENNESSEE KNOXVILLE

Subgraph Matching



37

THE UNIVERSITY of TENNESSEE KNOXVILLE

All Ranking Functions

Edge Based

- Absolute Difference
- Percent Difference
- Euclidean Distance
- Euclidean Distance, threshold based
- Euclidean Distance, inverse sum

Node Based

- Intersection
- Graph Matching
- Hamming Distance

Ensemble Style

- BCS
- Olympic Score

38

THE UNIVERSITY of TENNESSEE KNOXVILLE

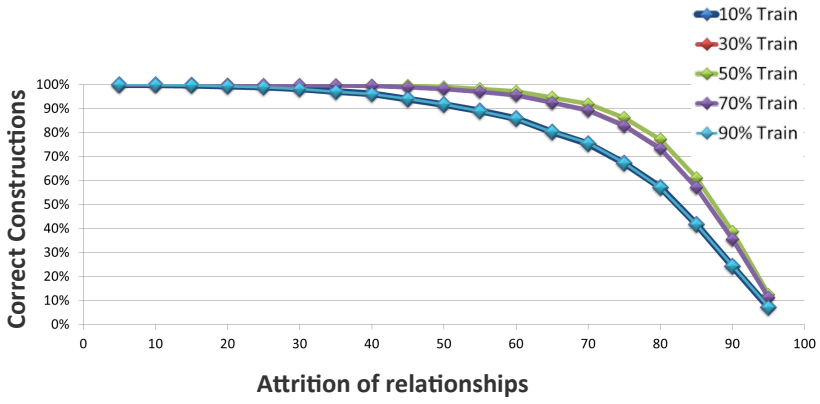
Graph Prints: Test Cases

- Number of People: 1K, 10K, 100K
- Length of time: 6
- Attrition on relationships: varies
- Training/Testing Split: varies
- Model Types: Binary and Weighted
- Total trials: 57,000 example graphs

39

THE UNIVERSITY of TENNESSEE KNOXVILLE

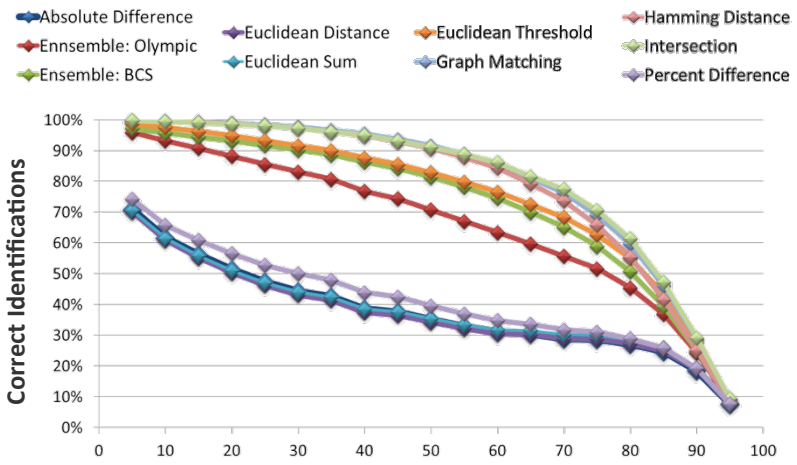
Validation of Approach



40

THE UNIVERSITY of TENNESSEE KNOXVILLE

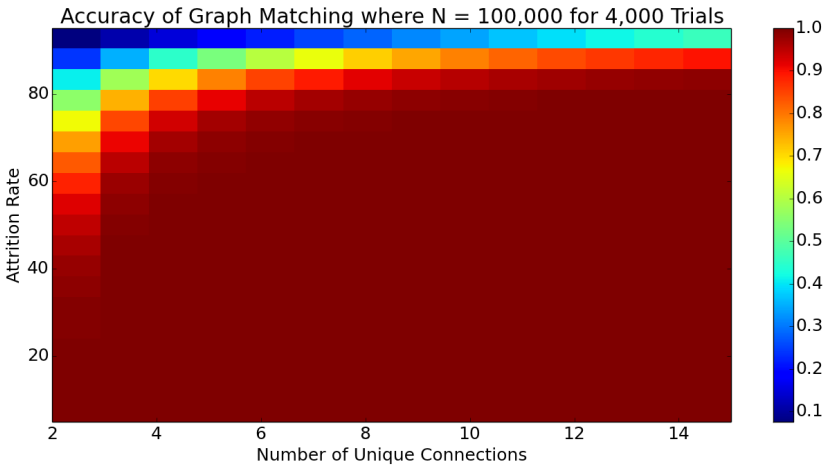
Ranking Function Performance



41

THE UNIVERSITY of TENNESSEE KNOXVILLE

Best of the Best



42

THE UNIVERSITY of TENNESSEE KNOXVILLE

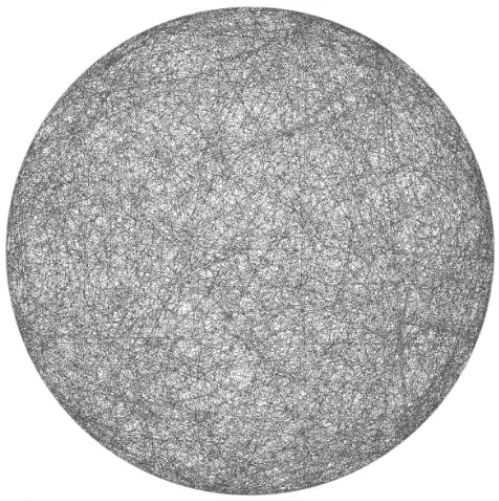
Social Fingerprinting:

Picking the individual out of crowd

43

THE UNIVERSITY of TENNESSEE KNOXVILLE

What does this look like?



44

THE UNIVERSITY of TENNESSEE KNOXVILLE

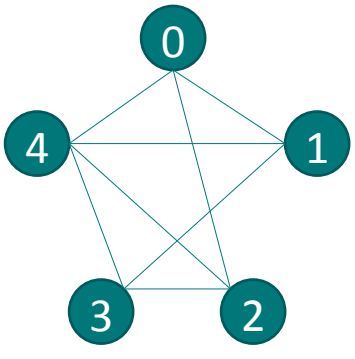
Social Fingerprint Procedure:

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

45

THE UNIVERSITY of TENNESSEE KNOXVILLE

Construction:



Time *t*

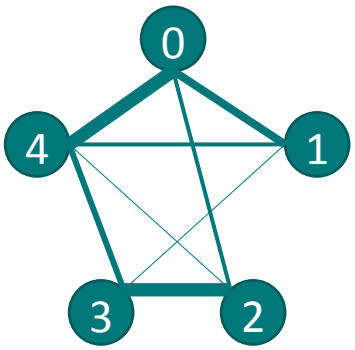
$$\begin{array}{c} \\ p_0 \\ p_1 \\ p_2 \\ p_3 \\ p_4 \end{array} \begin{array}{c} \begin{array}{ccccc} p_0 & p_1 & p_2 & p_3 & p_4 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1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

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

46

THE UNIVERSITY of TENNESSEE KNOXVILLE

Construction:



Time *t*

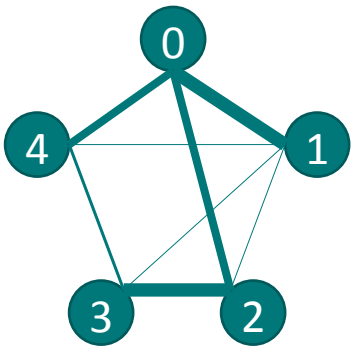
$$\begin{array}{c} \\ p_0 \\ p_1 \\ p_2 \\ p_3 \\ p_4 \end{array} \begin{array}{c} \begin{array}{ccccc} p_0 & p_1 & p_2 & p_3 & p_4 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \end{array} \\ \begin{bmatrix} 0 & 5 & 2 & 0 & 9 \\ 5 & 0 & 0 & 1 & 2 \\ 2 & 0 & 0 & 9 & 1 \\ 0 & 1 & 9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

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

47

THE UNIVERSITY of TENNESSEE KNOXVILLE

Construction: Query Vectors



Time *t + 1*

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

48

THE UNIVERSITY of TENNESSEE KNOXVILLE

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

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

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

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

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

49

THE UNIVERSITY of TENNESSEE KNOXVILLE

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

	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**

50

THE UNIVERSITY of TENNESSEE KNOXVILLE

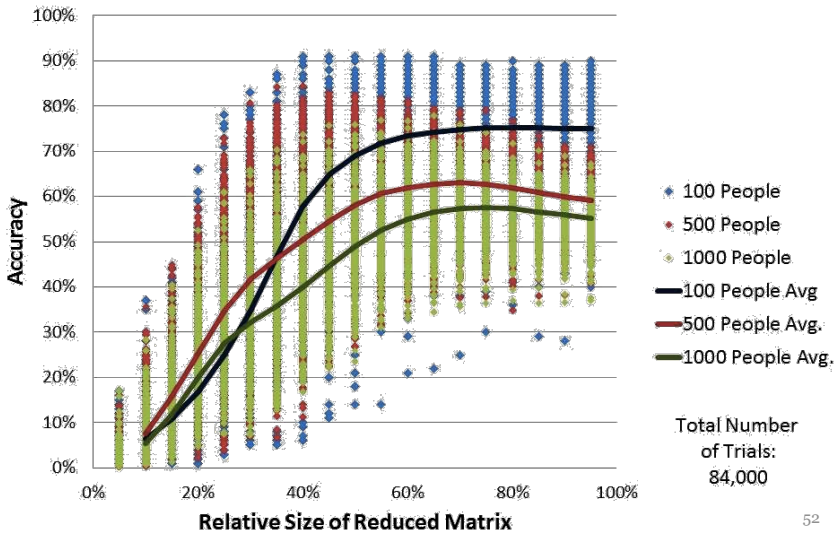
SDD: Test Cases

- Number of People: 100, 500 or 1,000
- Length of time: 12
- Attrition on relationships: 37%
- Low Rank Approximation rank: varies
- Training/Testing Split: varies
- Total trials: 84,000

51

THE UNIVERSITY of TENNESSEE KNOXVILLE

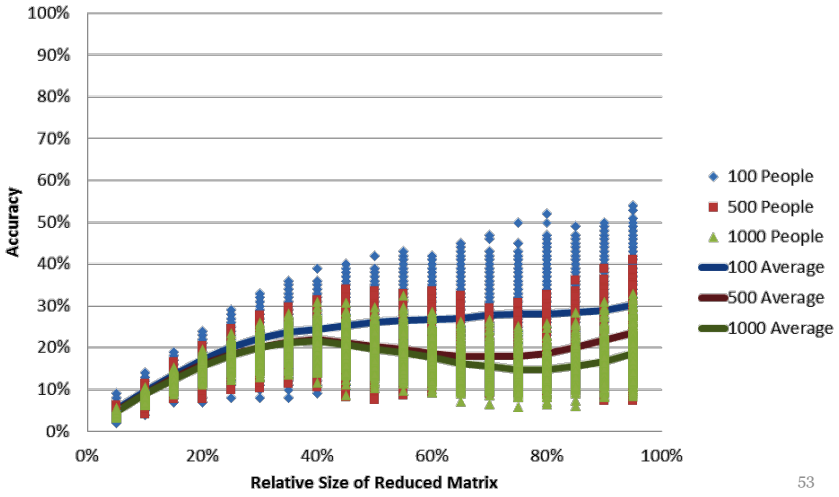
Binary Model Results



52

THE UNIVERSITY of TENNESSEE KNOXVILLE

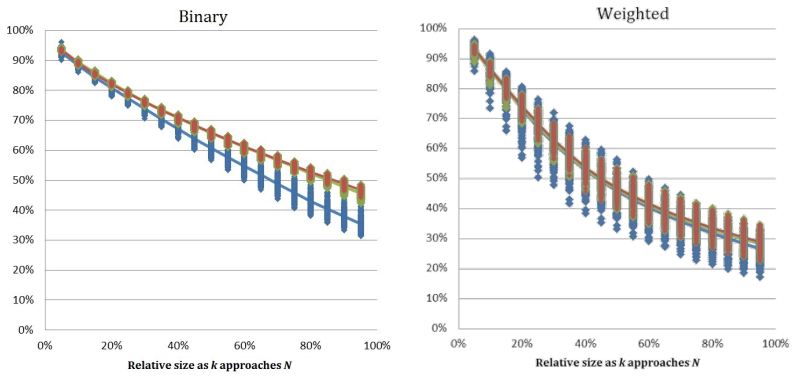
Weighted Model Results



53

THE UNIVERSITY of TENNESSEE KNOXVILLE

Binary vs Weighted Model Error



54

THE UNIVERSITY of TENNESSEE KNOXVILLE

SDD: Case Study

- Number of People: 100
- Length of time: 12
- Attrition on relationships: 37%
- Low Rank Approximation rank: 75%
- Training/Testing Split: 50/50

55

THE UNIVERSITY of TENNESSEE KNOXVILLE

SDD: Case Study Results

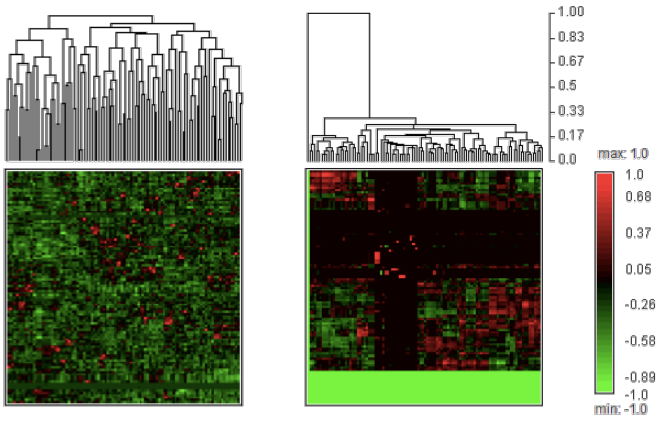
- Binary Model Accuracy: 81%
- Weighted Model Accuracy: 27%

	WM: Incorrect	WM: Correct
BM: Incorrect	12	7
BM: Correct	61	20

56

THE UNIVERSITY of TENNESSEE KNOXVILLE

Model Insights



Binary Model Similarity Matrix

Weighted Model Similarity Matrix

THE UNIVERSITY of TENNESSEE KNOXVILLE

Accomplishments and Contributions

- Introduced a novel problem
- Created and released a modeling environment: SOFAS
- Modeled and defined first approaches for social fingerprint identification

THE UNIVERSITY of TENNESSEE KNOXVILLE

Open Research

- New models for social fingerprint detection
- Additional features and graph statistics of the SOFA software
- Fingerprint Evasion

59

THE UNIVERSITY of TENNESSEE KNOXVILLE

Acknowledgements

Dr. Michael Berry

Dr. Judy Day

Dr. Jens Gregor

Dr. Bruce MacLennan

NSF

SCALE-IT
Scalable Computing and Leading Edge Innovative Technologies

Link Analytics

CISML

pokitdok

60

THE UNIVERSITY of TENNESSEE KNOXVILLE

Social Fingerprinting:
Identifying Users of Social Networks by their Data Footprint

The University of Tennessee
College of Engineering Dissertation Defense
Dr. Denise Koessler Gosnell, PhD Computer Science
(Nerdery)