Tutorial

Data-driven **Behavioral Modeling** of Mobile Users for Analysis, Simulation and Design of Future Mobile Social Networks

Ahmed Helmy

Computer and Information Science and Engineering (CISE) Department

University of Florida

helmy@ufl.edu , *http://www.cise.ufl.edu/~helmy*

Founder & Director: Wireless Mobile Networking Lab *http://nile.cise.ufl.edu*

Funded by:

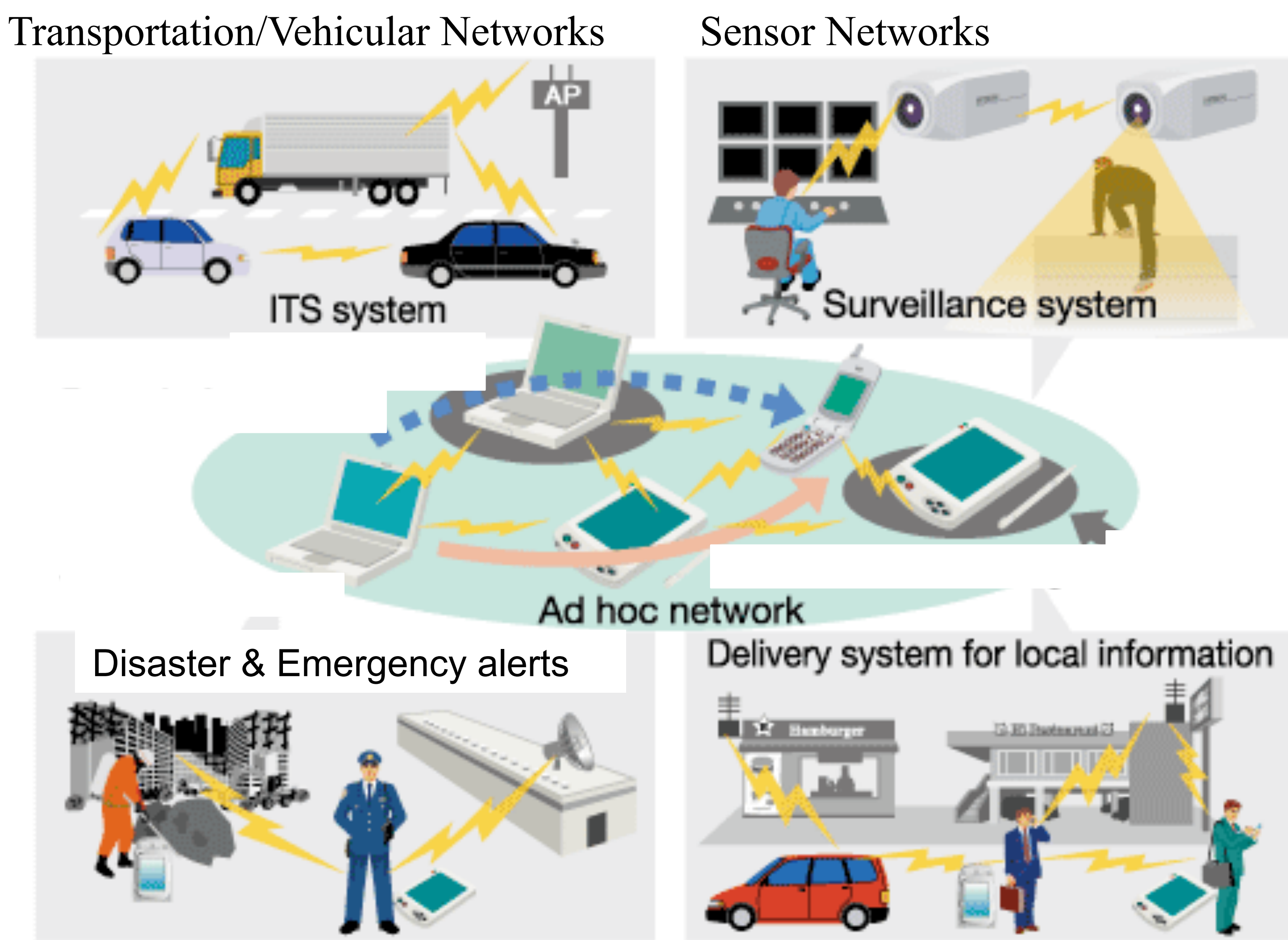
Outline

- Mobile Ad Hoc Networks & Delay Tolerant Networks (intro)
  - Proliferation of mobile devices, tight user-device coupling
  - Opportunities, potential behavior-aware applications
- The paradigm shift in modeling and design
  - Data-driven approach
  - The *TRACE* framework
- Tracing and measurements
  - Overview of *MobiLib* and *Crawdad* community libraries
  - Mining and analyzing the traces, tool and method discussion
- Individual behavioral modeling
  - Location preferences, periodic re-appearance
  - Time variant community (*TVC*) model

Outline (contd.)

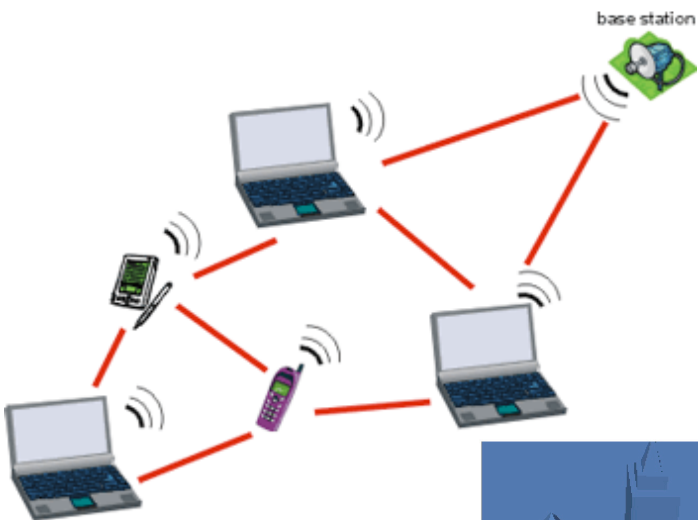
- Pair-wise (encounter) modeling
  - Encounter graphs
  - Small world analysis
- Collective behavior and clustering
  - Association matrix, and similarity of behavior
  - Clustering based on mobility preferences using WLAN MAC traps
  - Clustering based on web access patterns and interests using Netflow
  - Co-clustering, - Self-organizing maps
- Applications, protocols and services
  - Interest-aware, privacy preserving communications (profile-cast)
  - Behavior-based trust
  - Participatory sensing and crowd sourcing

Networked Mobile Societies Everywhere, Anytime

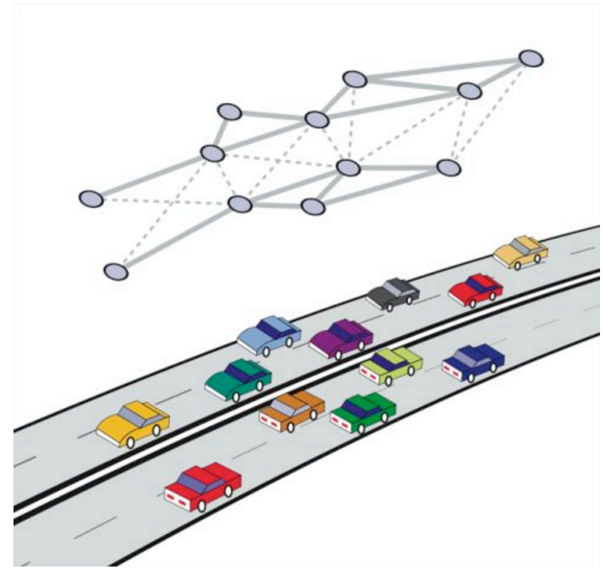


Mobile Ad hoc, Sensor and Delay Tolerant Networks

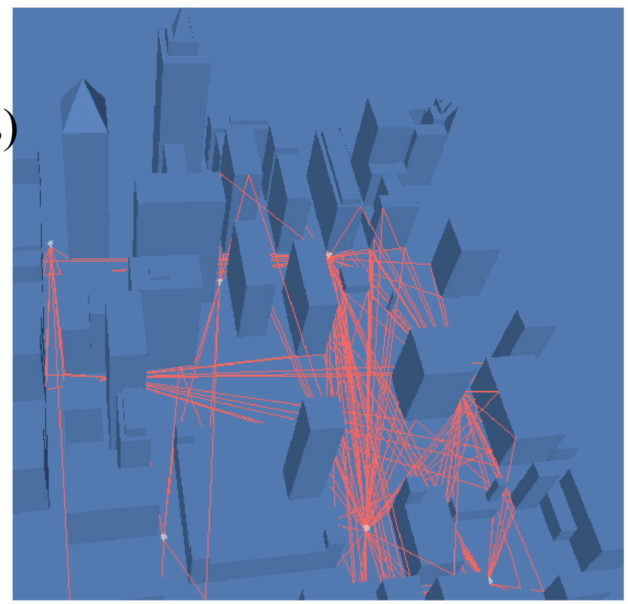
Example Ad hoc Networks & DTNs



Mobile devices (laptop, PDAs)



Vehicular Networks on Highways



Hybrid urban ad hoc network (vehicular, pedestrian, hot spots,…)

Wireless Mobile Ad hoc Networks (MANETs)

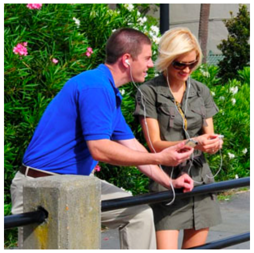
- A Mobile Ad hoc Network: collection of mobile devices forming a multi-hop wireless network with no infrastructure
- Ad hoc networks can be highly dynamic due to mobility, topology change, wireless characteristics, lack of infrastructure, limited node/device capabilities

Delay/Disruption Tolerant Networks (DTNs)

- Intermittently connected Ad hoc Network. Not all paths are valid at any given time, but over a time span
- Routing performed in time and space: forward, store, carry, forward, … challenging if mobility is not deterministic

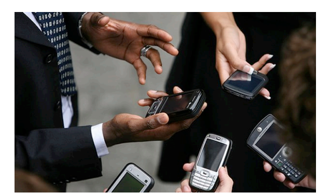
In MANETs and DTNs, cooperating mobile nodes 'are' the network

Emerging Behavior-Aware Services



- Tight coupling between users, devices
  - Devices can infer user preferences, behavior
  - Capabilities: comm, comp, storage, sensing
- New generation of behavior-aware protocols
  - Behavior: mobility, interest, trust, friendship,…
  - Apps: interest-cast, participatory sensing, crowd sourcing, mobile social nets, alert systems, …

New paradigms of communication?!



Paradigm Shift in Protocol Design

Used to:

Design general purpose protocols
Evaluate using models (random mobility, traffic, …)
Deployment context: Modify to improve performance and failures for specific context

– May end up with suboptimal performance or failures due to lack of context in the design

Propose to:

Analyze, model deployment ***context***
Design *'application class'-specific* parameterized protocols
Utilize insights from context analysis to fine-tune protocol parameters

Problem Statement

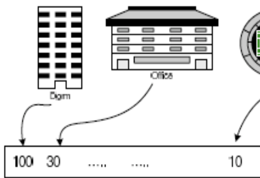
- How to gain insight into deployment context?
- How to utilize insight to design future services?

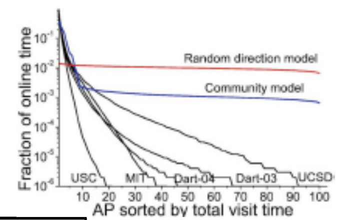
Approach

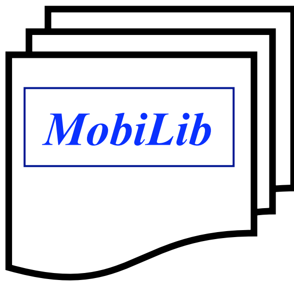
- Extensive trace-based analysis to identify dominant trends & characteristics
- Analyze user behavioral patterns
  - Individual user behavior and mobility
  - Collective user behavior: grouping, encounters
- Integrate findings in modeling and protocol design
  - I. User mobility modeling – II. Behavioral grouping
  - III. Information dissemination in mobile societies, *profile-cast*

The *TRACE* framework

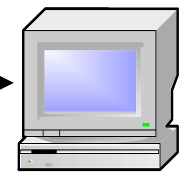
	Duration	Records	Total Users	Access points (or bldgs)
USC-WLAN	Dec 03-Jun 08	50 M	55,500	79 ports (03), 161 (08)
USC-DHCP	Dec 03-Jun 08	60 M	55,500	79 ports (03), 161 (08)
USC-netflow	Apr 05-Jun 08	50 B	50,000	161 ports
UF-WLAN	Jun 07-Current	45 M	105,500	784 Access points
UF-DHCP	Jun 07-Current	10 M	105,500	784 Access points
UF-netflow	to start Sep 09	n/a	n/a	784+ Access points







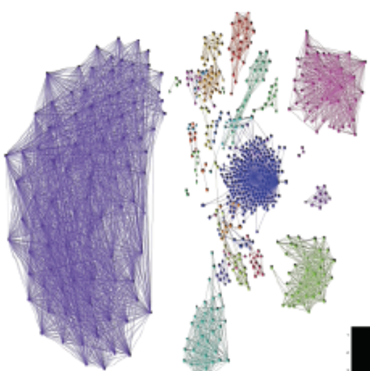
$$\begin{pmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{t,1} & \cdots & x_{t,n} \end{pmatrix}$$

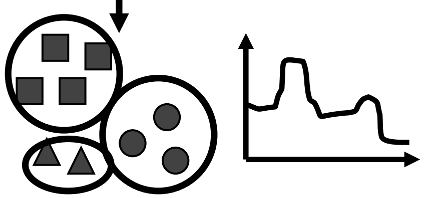


Trace

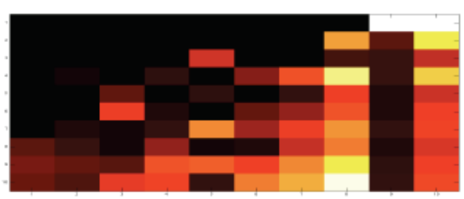
Represent

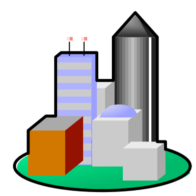
Analyze

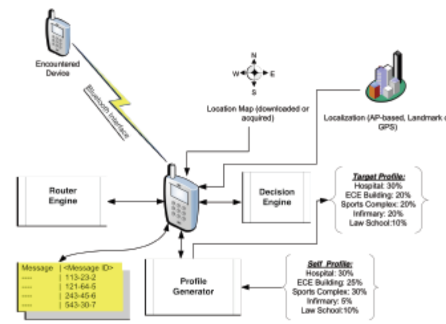




Characterize, ***C***luster

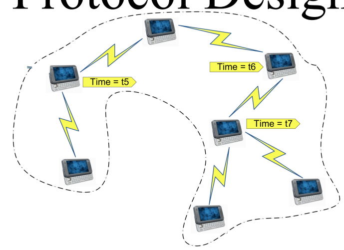






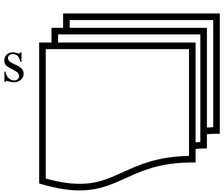
Employ

(Modeling, Protocol Design)



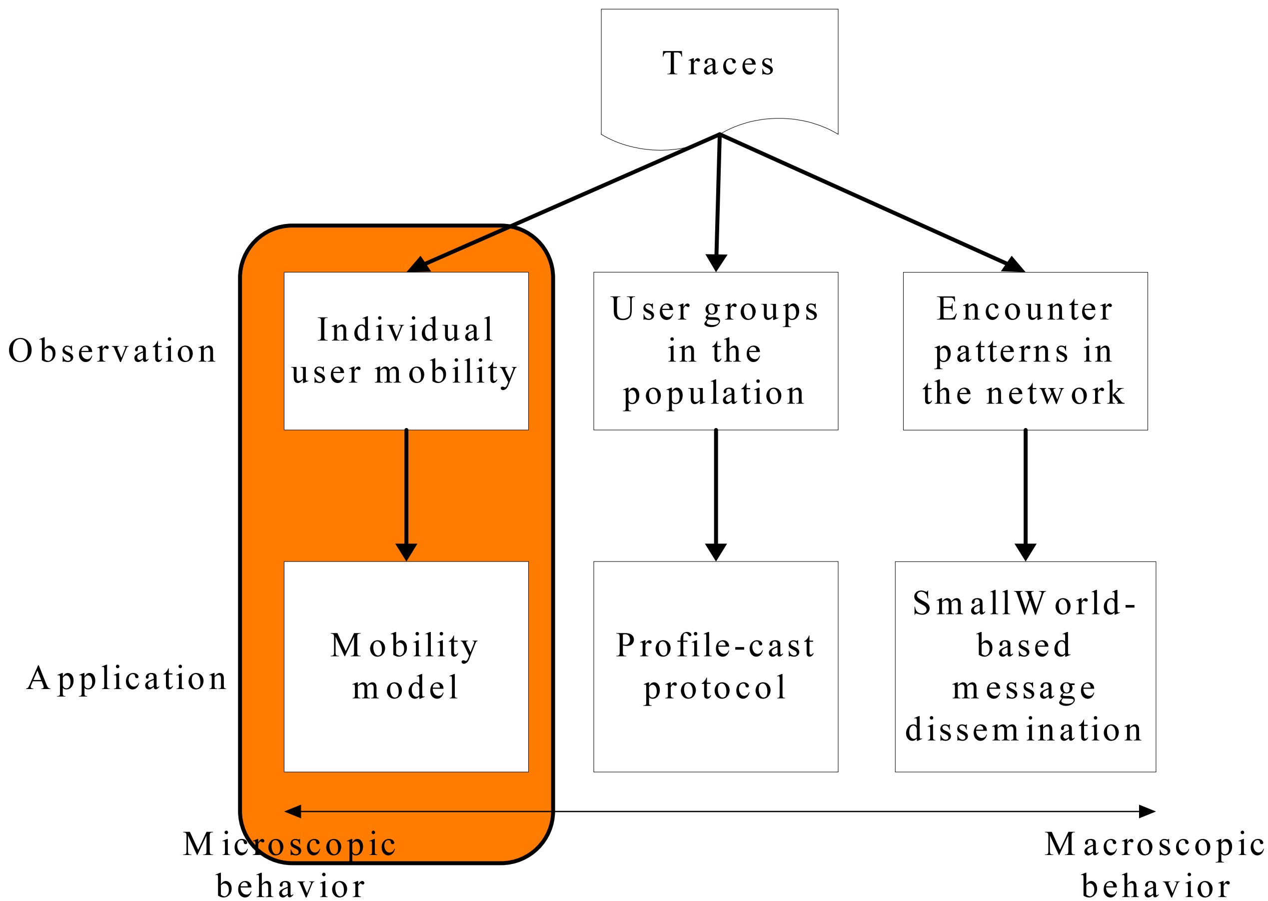
Community-wide Wireless/Mobility Library

- Library of
  - Measurements from Universities, vehicular networks
  - *Realistic* models of behavior (mobility, traffic, encounters)
  - Simulation benchmarks - Tools for trace data mining
- Available libraries:
  - CRAWDAD (Dartmouth, '05-) *crawdad.cs.dartmouth.edu*
    MobiLib (USC & UFL, '04-) *nile.cise.ufl.edu/MobiLib*
    - 60+ Traces from: *USC, Dartmouth, MIT, UCSD, UCSB, UNC, UMass, GATech, Cambridge, UFL*, … (tens of millions of traces)
    - Tools for mobility modeling (*IMPORTANT*, *TVC*), data mining
- Types of traces:
  - Campuses (WLANs), Conference AP and encounter traces
  - Municipal (off-campus) wireless APs
  - GPS logs for taxi cabs, buses

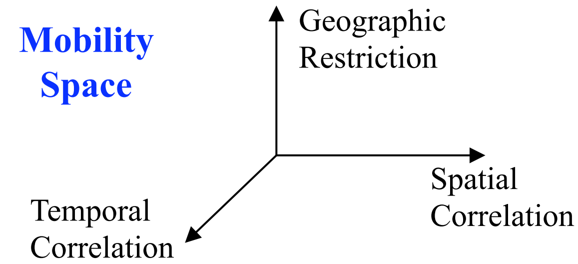


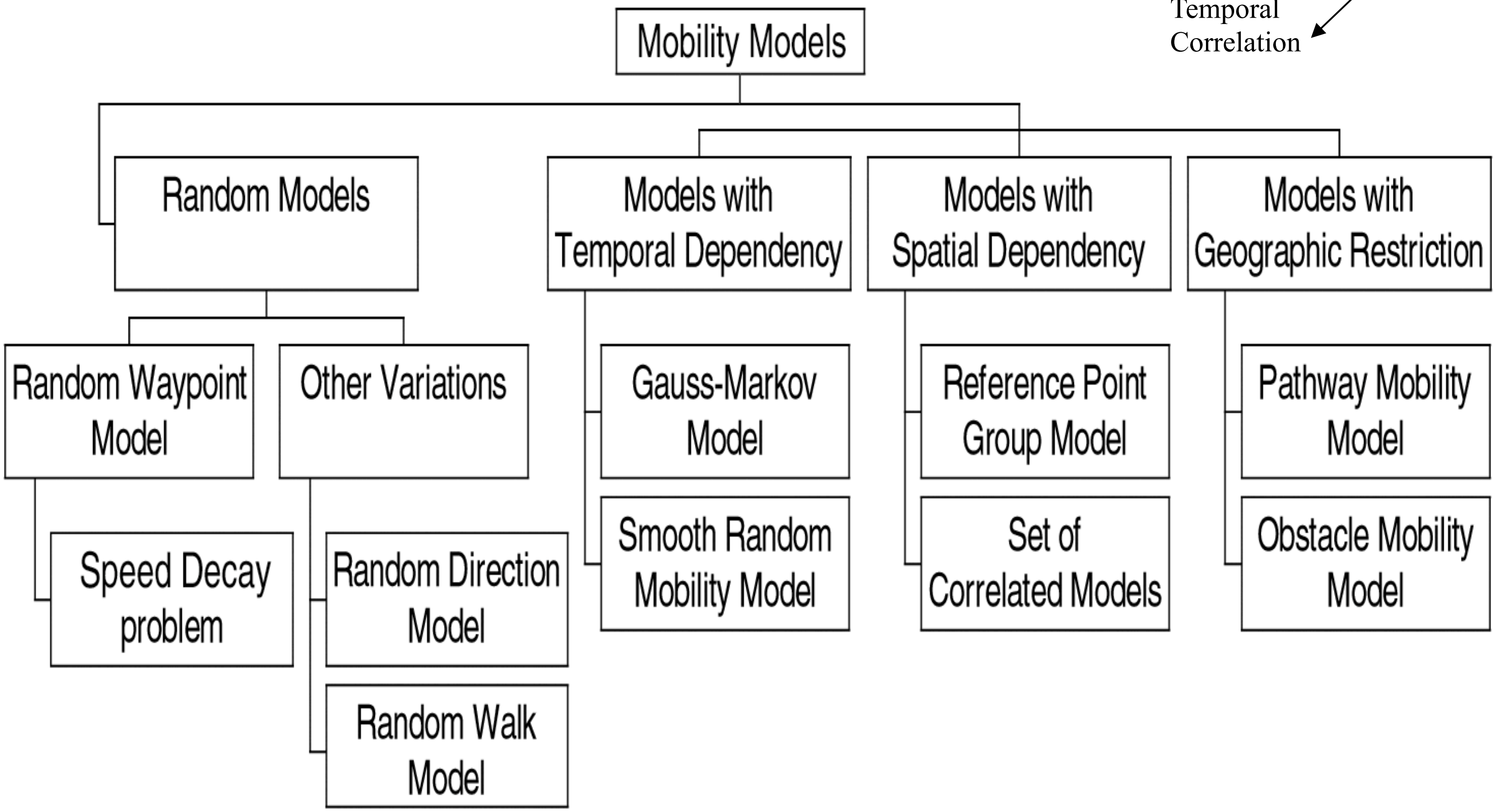
***T*race**

Case study I – Individual Mobility



Classification of Mobility Models

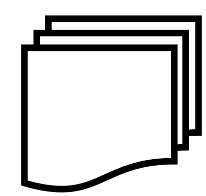




* F. Bai, A. Helmy, "A Survey of Mobility Modeling and Analysis in Wireles Adhoc Networks", Book Chapter in the book "Wireless Ad Hoc and Sensor Networks", Kluwer Academic Publishers, June 2004.

Wireless Networks and Mobility Measurements

- In our case studies we use WLAN traces
  - From University campuses & corporate networks (4 universities, 1 corporate network)
  - The largest data sets about wireless network users available to date (# users / lengths)
  - No bias: not "special-purpose", data from all users in the network
- We also analyze
  - Vehicular movement trace (Cab-spotting)
  - Human encounter trace (at Infocom Conf)



Trace

IMPACT: Investigation of Mobile-user Patterns Across University Campuses using WLAN Trace Analysis*

- 4 major campuses – 30 day traces studied from 2+ years of traces
- Total users > 12,000 users - Total Access Points > 1,300

Trace source	Trace duration	User type	Environment	Collection method	Analyzed part
MIT	7/20/02 – 8/17/02	Generic	3 corporate buildings	Polling	Whole trace
Dartmouth	4/01/01 – 6/30/04	Generic w/ subgroup	University campus	Event-based	July ’03 April ’04
UCSD	9/22/02 – 12/8/02	PDA only	University campus	Polling	09/22/02-10/21/02
USC	4/20/05 – 3/31/06	Generic	University campus	Event-based (Bldg)	04/20/05-05/19/05

* W. Hsu, A. Helmy, “***IMPACT***: Investigation of Mobile-user Patterns Across University Campuses using WLAN Trace Analysis”, two papers at *IEEE Wireless Networks Measurements (WiNMee)*, April 2006 and *IEEE Transactions on Mobile Computing*, Nov 2010.

Case Study I: Goal

- To understand the mobility/usage pattern of individual wireless network users
- To observe how environments/user type/trace-collection techniques impact the observations
- To propose a realistic mobility model based on empirical observations
  - That is mathematically tractable
  - That is capable of characterizing multiple classes of mobility scenarios

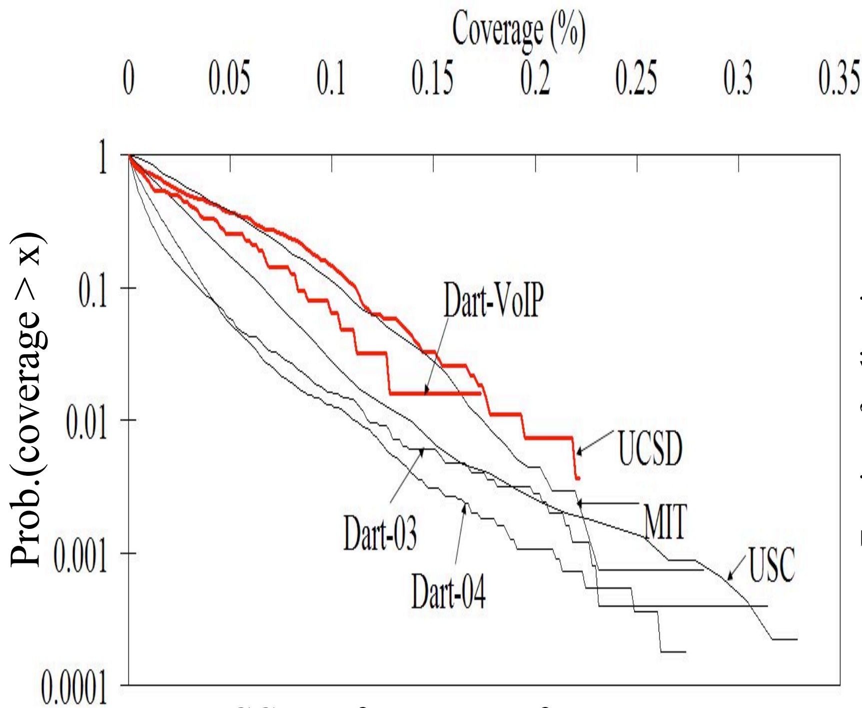
Metrics for Individual Mobility Analysis

- What kind of spatial preference do users exhibit?
  - The percentile of time spent at the most frequently visited locations
- What kind of temporal repetition do users exhibit?
  - The probability of re-appearance
- How often are the nodes present?
  - Percentage of "online" time

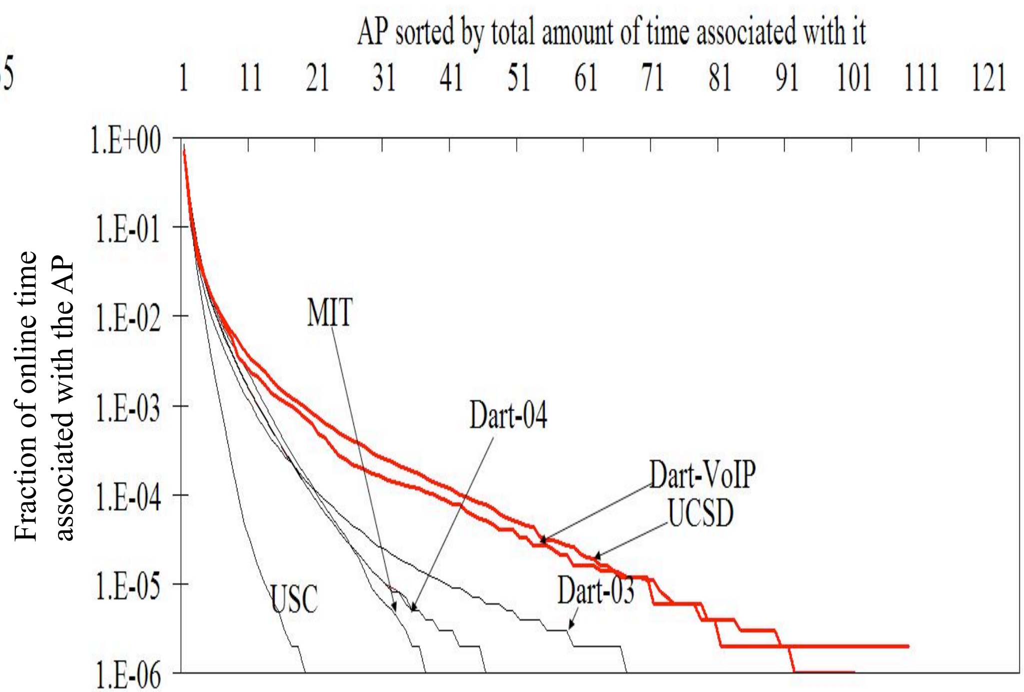
$$\begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{t,1} & \cdots & x_{t,n} \end{bmatrix}$$

***R**epresent*

Observations: Visited Access Points (APs)



CCDF of coverage of users
[percentage of visited APs]



Average fraction of time a MN associates with APs

- ***Individual users access only a very small portion of APs in the network.***
- ***On average a user spends more than 95% of time at its top 5 most visited APs.***
- ***Long-term mobility is highly skewed in terms of time associated with each AP.***
- ***Users exhibit "on"/"off" behavior that needs to be modeled.***

Repetitive Behavior

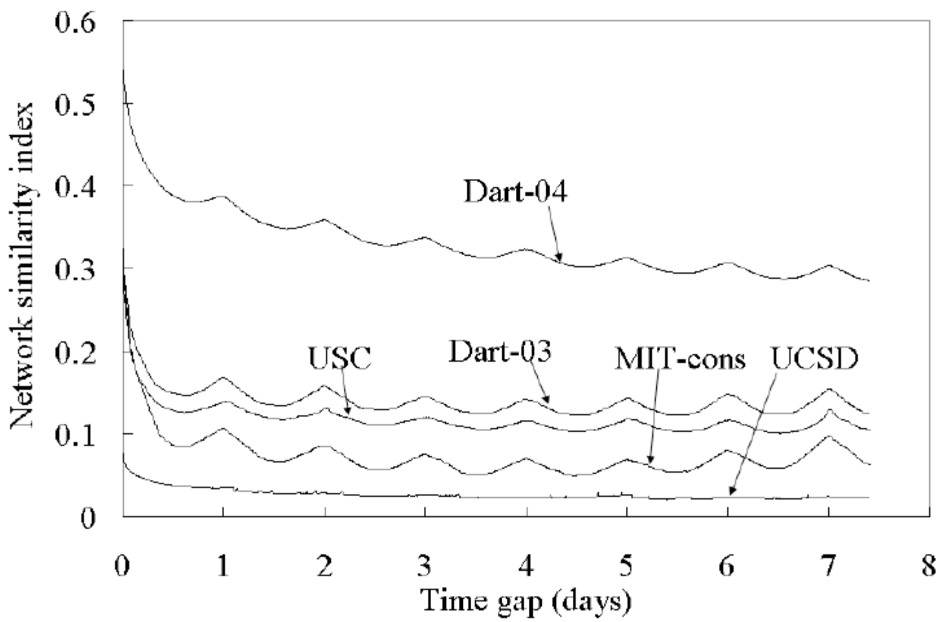


Fig. 7. Network similarity indexes. The peaks represent intervals for which there is high similarity.

- ***Clear repetitive patterns of association in wireless network users.***
- ***Typically, user association patterns show the strongest repetitive pattern at time gap of one day/one week.***

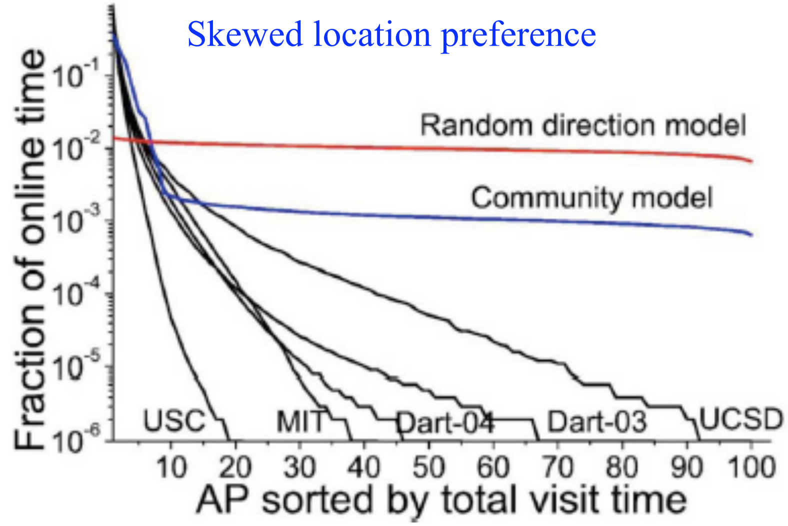
Mobility Models

- Mobility models are of crucial importance for the evaluation of wireless mobile networks [IMP03]*
- Requirements for mobility models
  - Realism (detailed behavior from traces)
  - Parameterized, tunable behavior
  - Mathematical tractability
- Related work on mobility modeling
  - Random walk/waypoint/dir models: mostly not realistic
  - Improved synthetic models (pathway, RPGM, WWP, FWY, MH) – more realistic, difficult to analyze, not repetitive
  - Trace-based model (T/T++): trace-specific, not general

* F. Bai, N. Sadagopan, A. Helmy, “IMPORTANT”, Infocom 2003, J. Adhoc Networks – El Sevier 2003

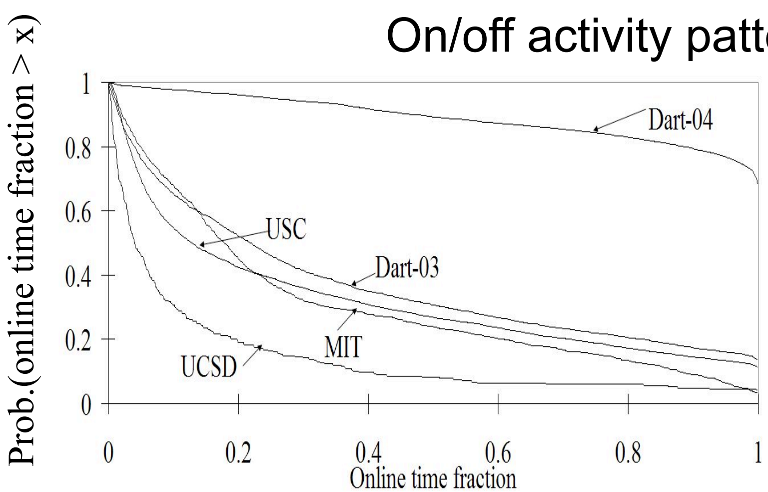
Spatio-temporal Mobility in WLANs

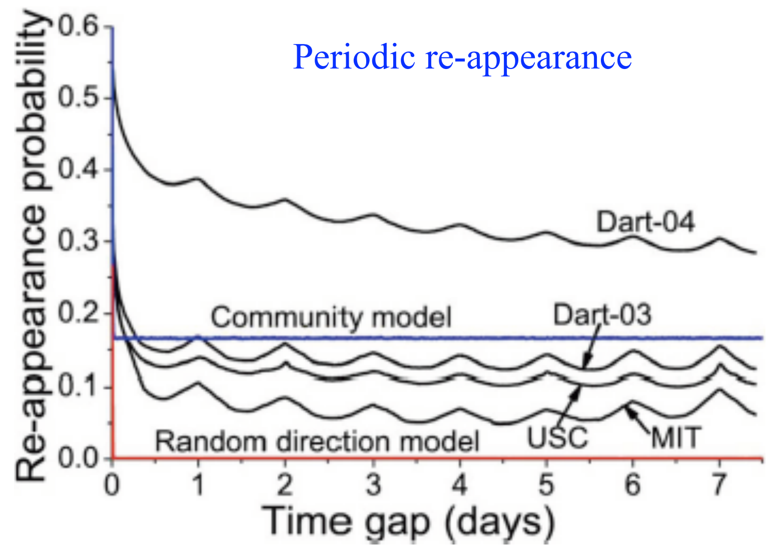
- Simple existing models are very different from the spatio-temporal characteristics in WLANs



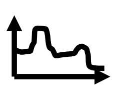
95% on-line time at 5 most visited APs

On/off activity pattern





Periodic repetition peaks daily/weekly

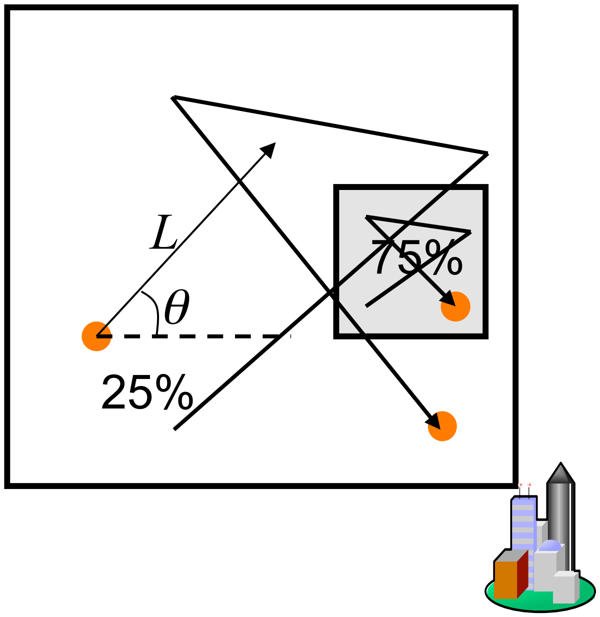


*C*haracterize

Time-variant Community (*TVC*) Model

(W. Hsu, Thyro, K. Psounis, A. Helmy, “Modeling Time-variant User Mobility in Wireless Mobile Networks”, IEEE INFOCOM, 2007, IEEE/ACM Transactions on Networking 2009)

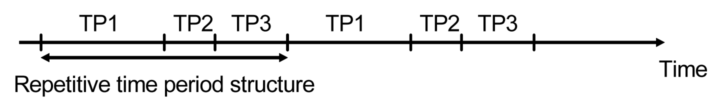
- Skewed location visiting preference
  - Create “communities” to be the preferred area of movement
  - Each node can have its own community
- Node moves with two different epoch types – Local or roaming
  - Each epoch is a random-direction, straight-line movement
  - Local epochs in the community
  - Roaming epochs around the whole simulation area



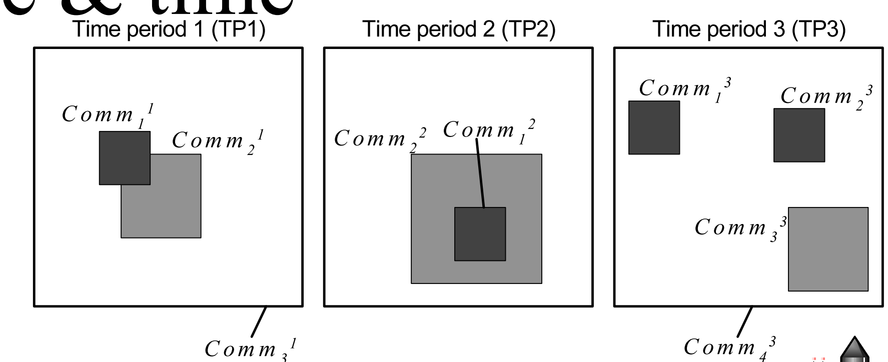
Employ

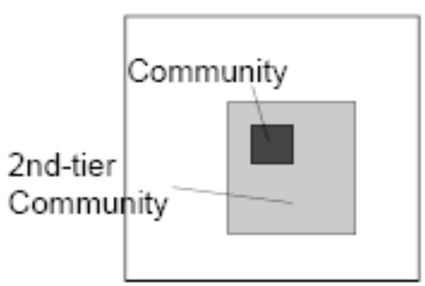
Tiered Time-variant Community (*TVC*) Model

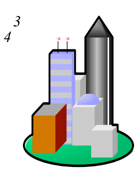
- Periodical re-appearance
  - Create structure in time – Periods
  - Node moves with different parameters in periods to capture time-dependent mobility
  - Repetitive structure



- Finer granularity in space & time
  - Multi-tier communities
  - Multiple time periods



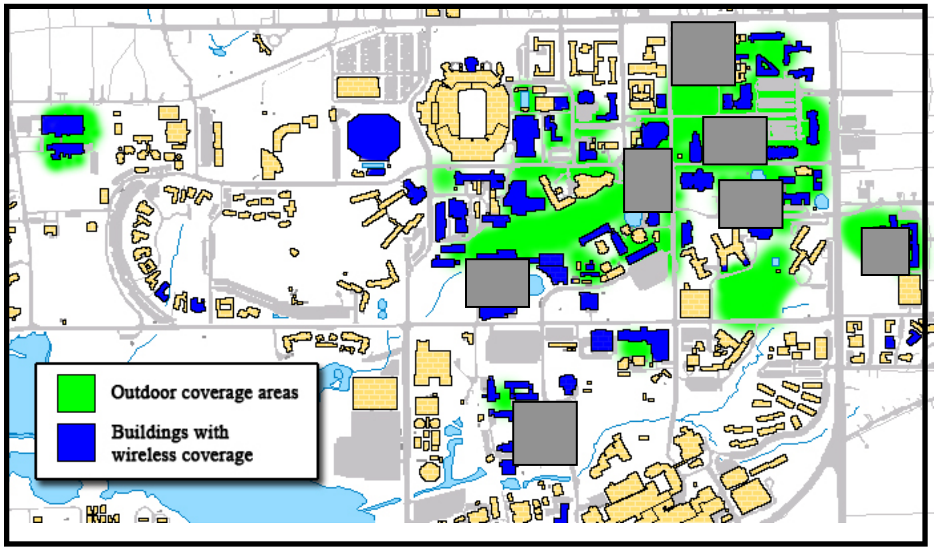




***E**mploy*

Using the TVC Model – Reproducing Mobility Characteristics

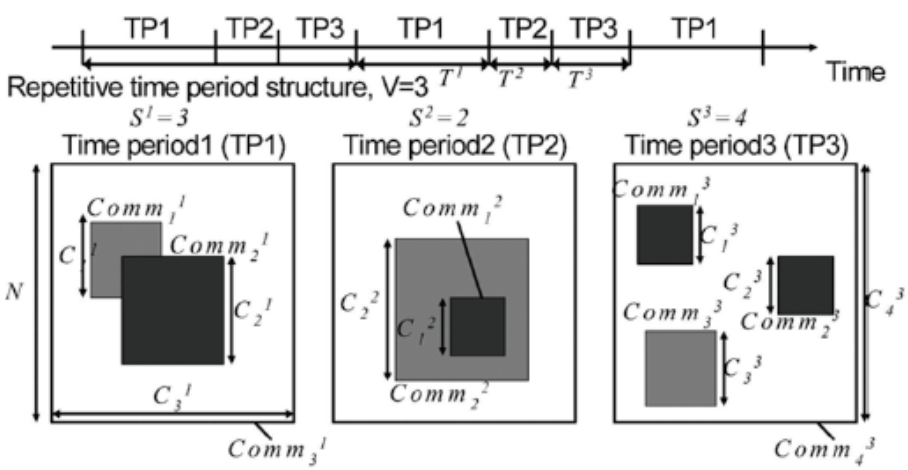
- (STEP1) Identify the popular locations; assign communities
- (STEP2) Assign parameters to the communities according to stats



- (STEP3) Add user on-off patterns (e.g., in WLAN users are usually ‘off’ when moving, in Cab spotting: vehicles ‘on’ when moving)

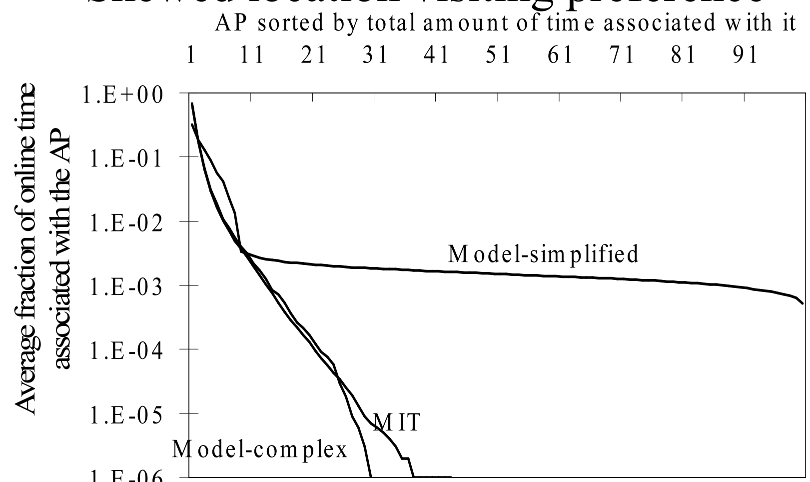
The *TVC* Model: Reproducing Mobility Characteristics

Time-Variant Community (***TVC***) Model:
1- Assigns communities (locations) to users to re-produce location visiting preference
2- Varies temporal assignment of communities to re-produce the periodic re-appearance



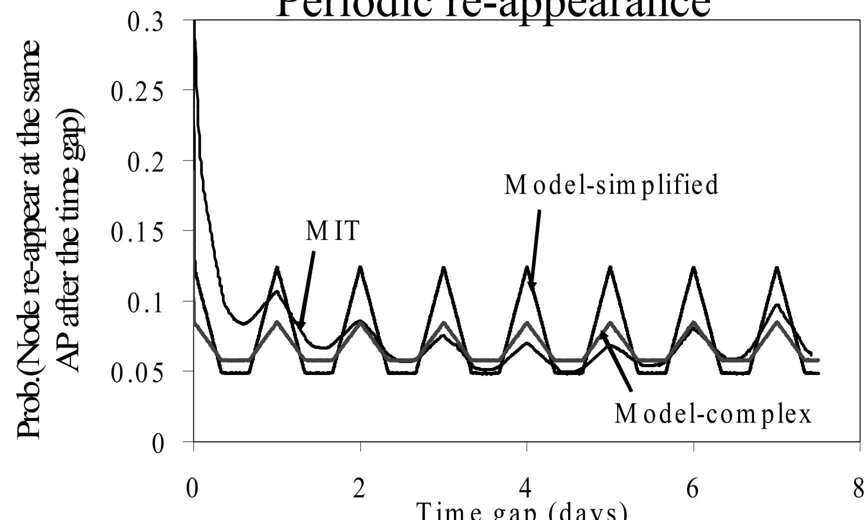
Time-variant mobility model, with three time periods and different numbers of communities in each period.

Skewed location visiting preference



CCDF

Periodic re-appearance

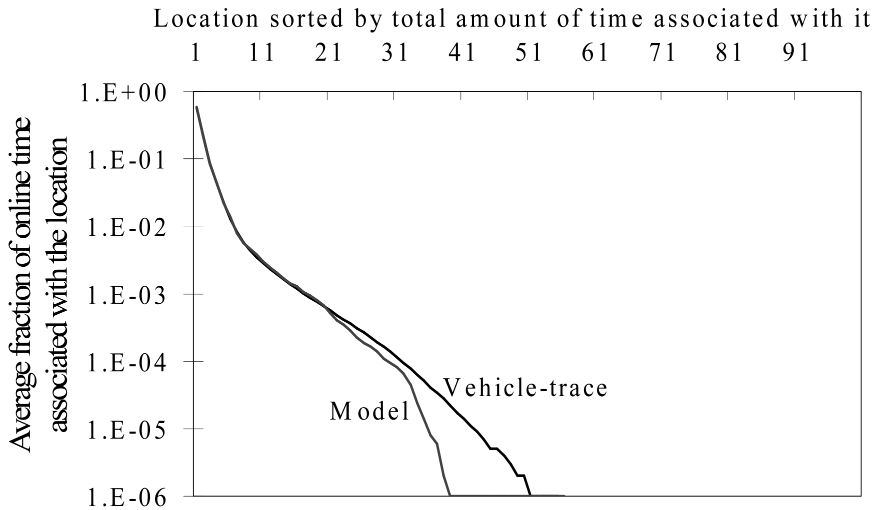


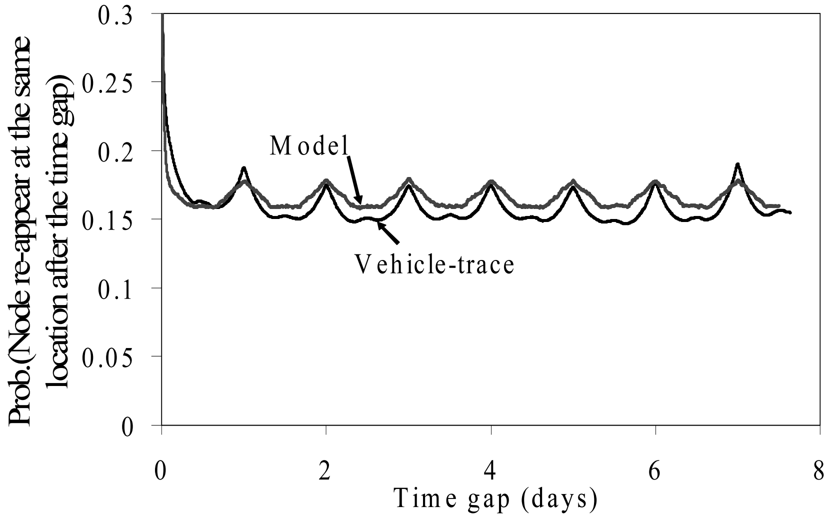
IEEE INFOCOM 2007
IEEE/ACM Trans. on Networking 2009

* Model-simplified: single community per node. Model-complex: multiple communities
** Similar matches achieved for USC and Dartmouth traces

Using the TVC Model – Reproducing Mobility Characteristics

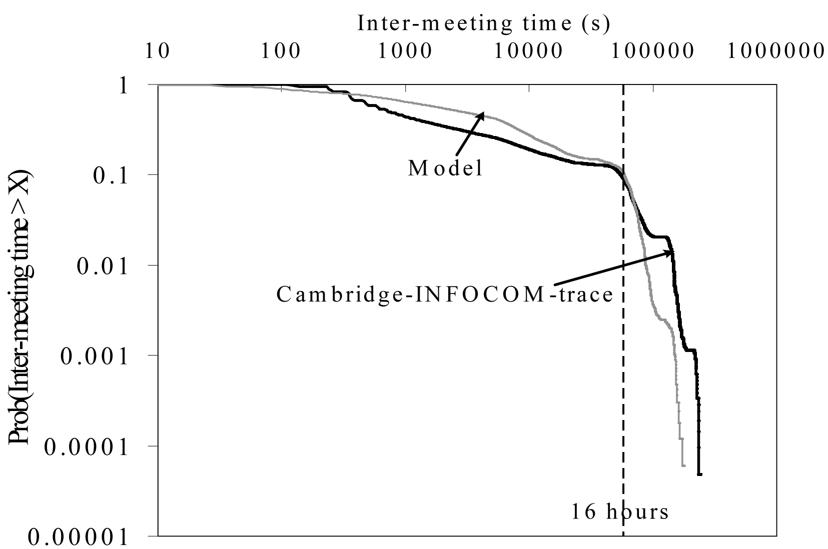
- Vehicular trace (Cab-spotting)



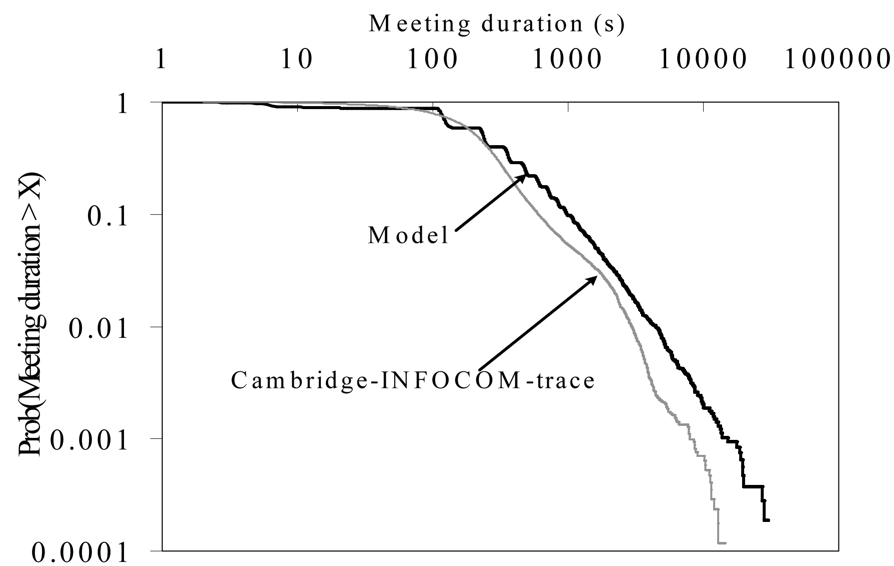


Using the *TVC* Model – Reproducing Mobility Characteristics

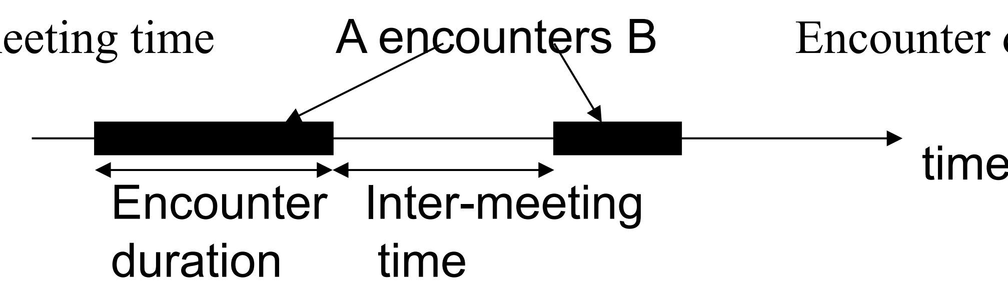
- Human encounter trace at a conference



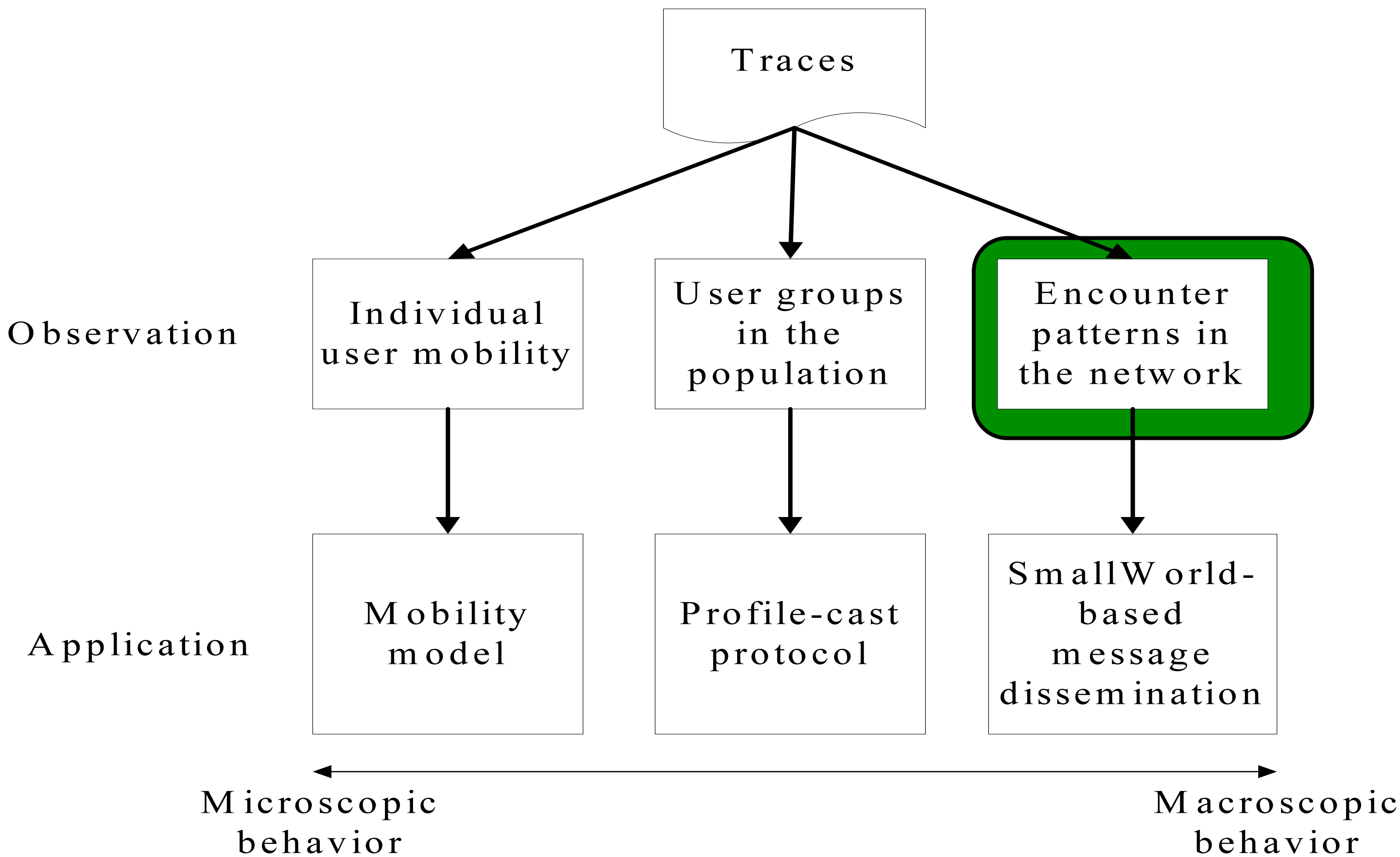
Inter-meeting time



Encounter duration



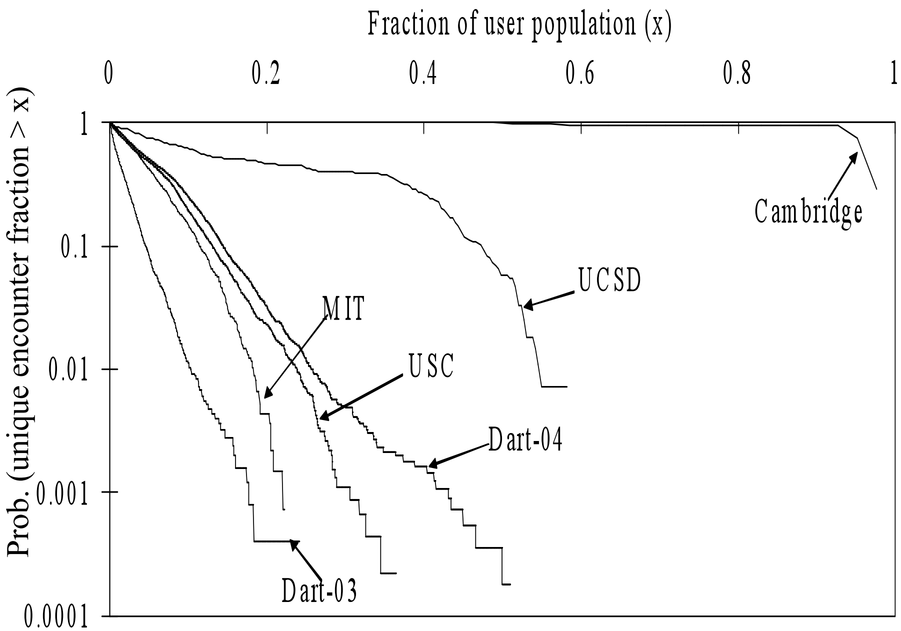
Case study II – Encounter Patterns



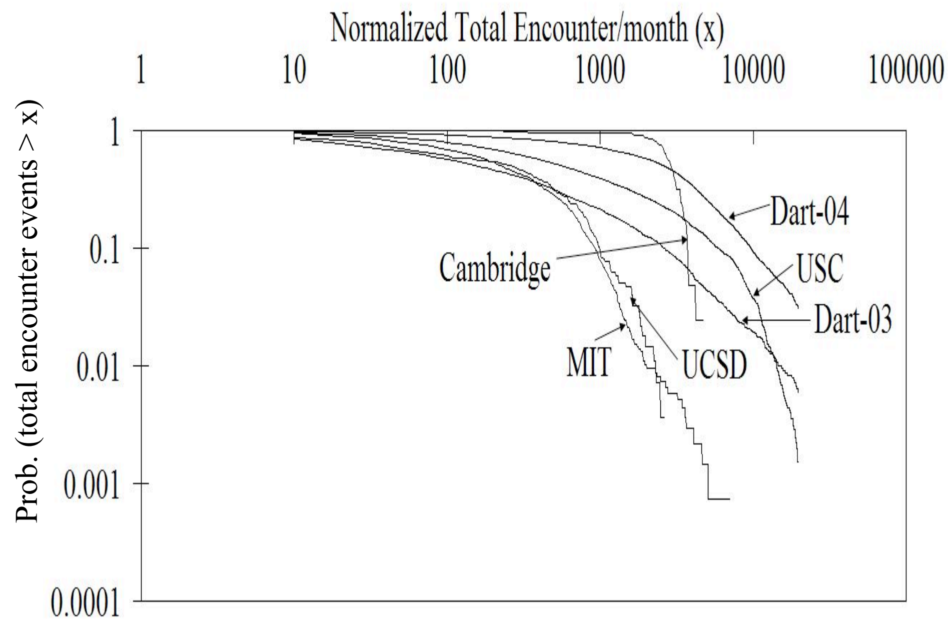
Case Study II: Goal

- Understand inter-node encounter patterns from a global perspective
  - How do we represent encounter patterns?
  - How do the encounter patterns influence network connectivity and communication protocols?
- Encounter definition:
  - In WLAN: When two mobile nodes access the same AP at the same time they have an 'encounter'
  - In DTN: When two mobile nodes move within communication range they have an 'encounter'

Observations: Nodal Encounters



CCDF of unique encounter count



CCDF of total encounter count

- ***In all the traces, the MNs encounter a small fraction of the user population.***
- ***A user encounters** 1.8%-6% **on average of the user population***
- ***The number of total encounters for the users follows a BiPareto distribution.***

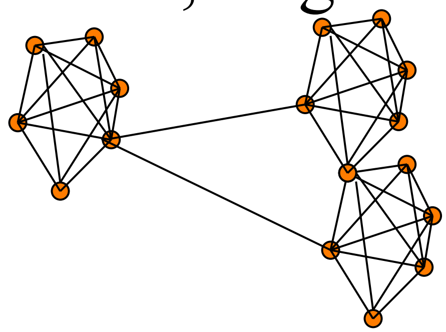
W. Hsu, A. Helmy, "On Nodal Encounter Patterns in Wireless LAN Traces", *IEEE Transactions on Mobile Computing (TMC)*, November 2010.

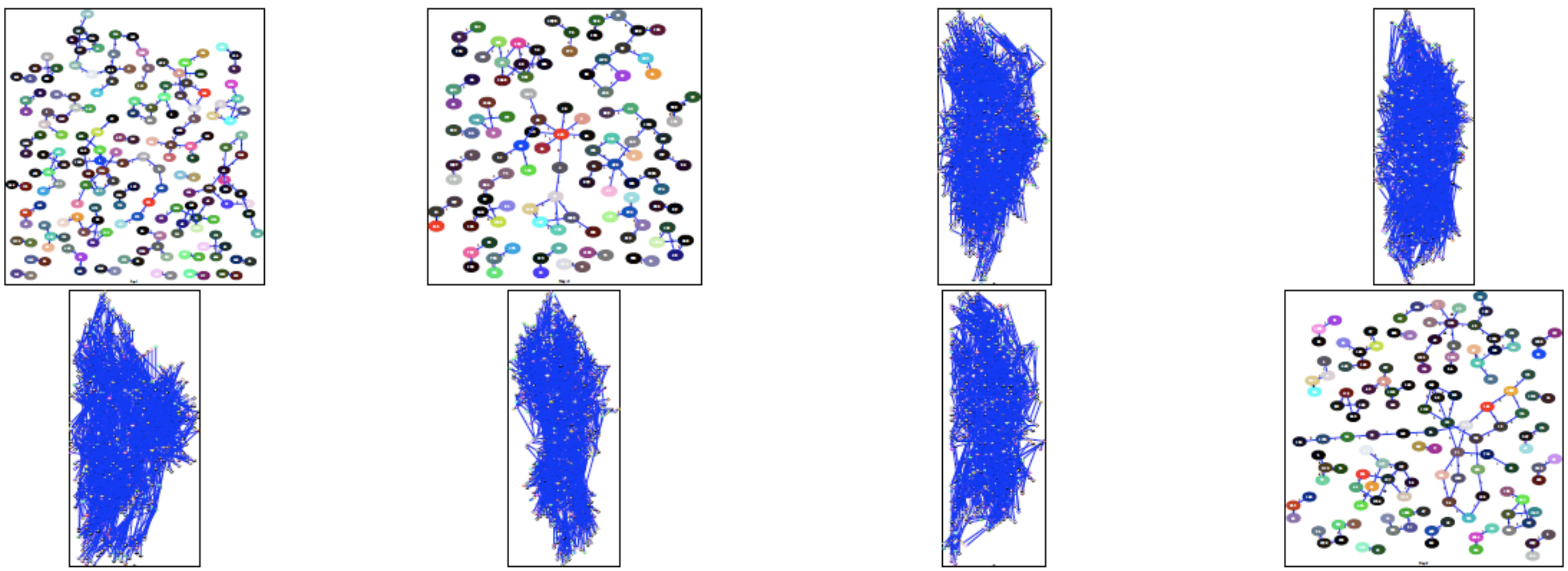
The Encounter graph

- Vertices: mobile nodes, Edges: node encounters

$$\begin{pmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{t,1} & \cdots & x_{t,n} \end{pmatrix}$$

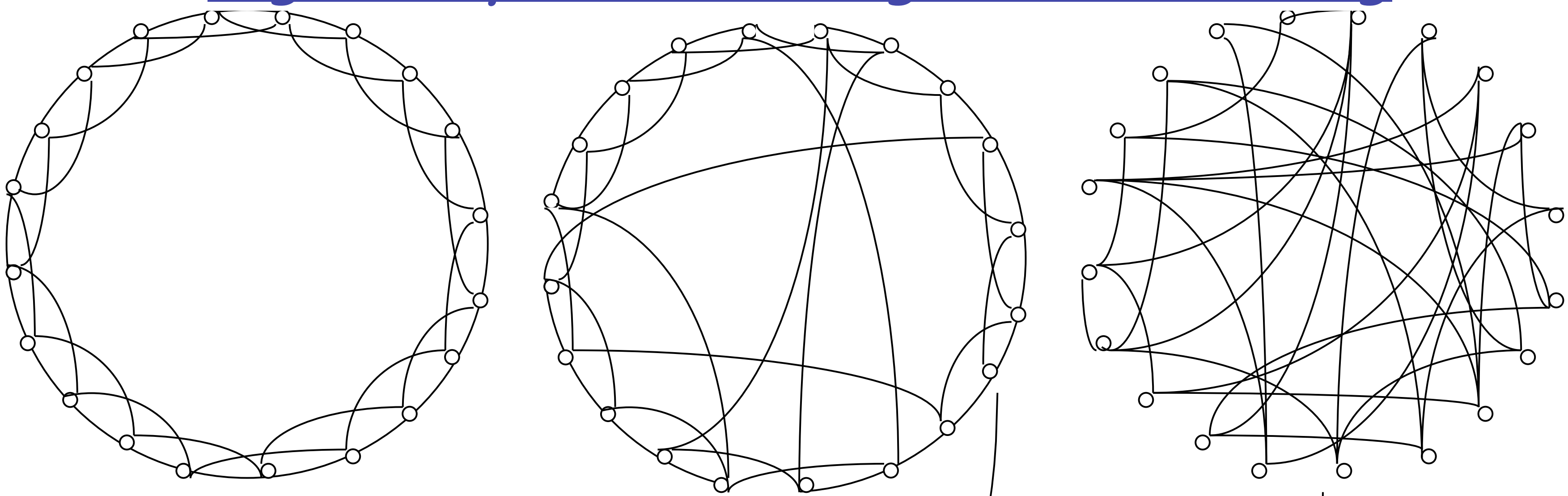
Represent





Daily encounter graphs for MIT trace

Bkgnd:Graphs , Path Length and Clustering



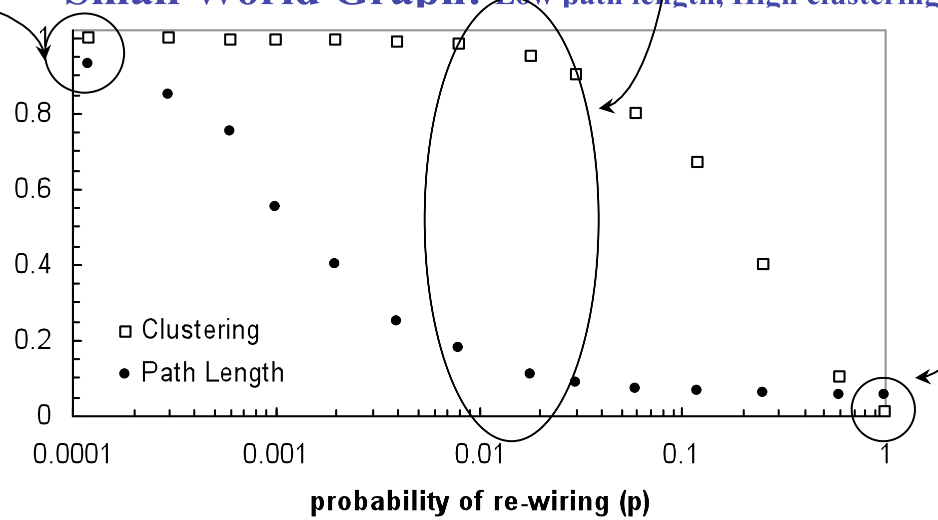
Regular Graph

- High path length
- High clustering

Small World Graph: **Low path length, High clustering**

Random Graph

- Low path length,
- Low clustering

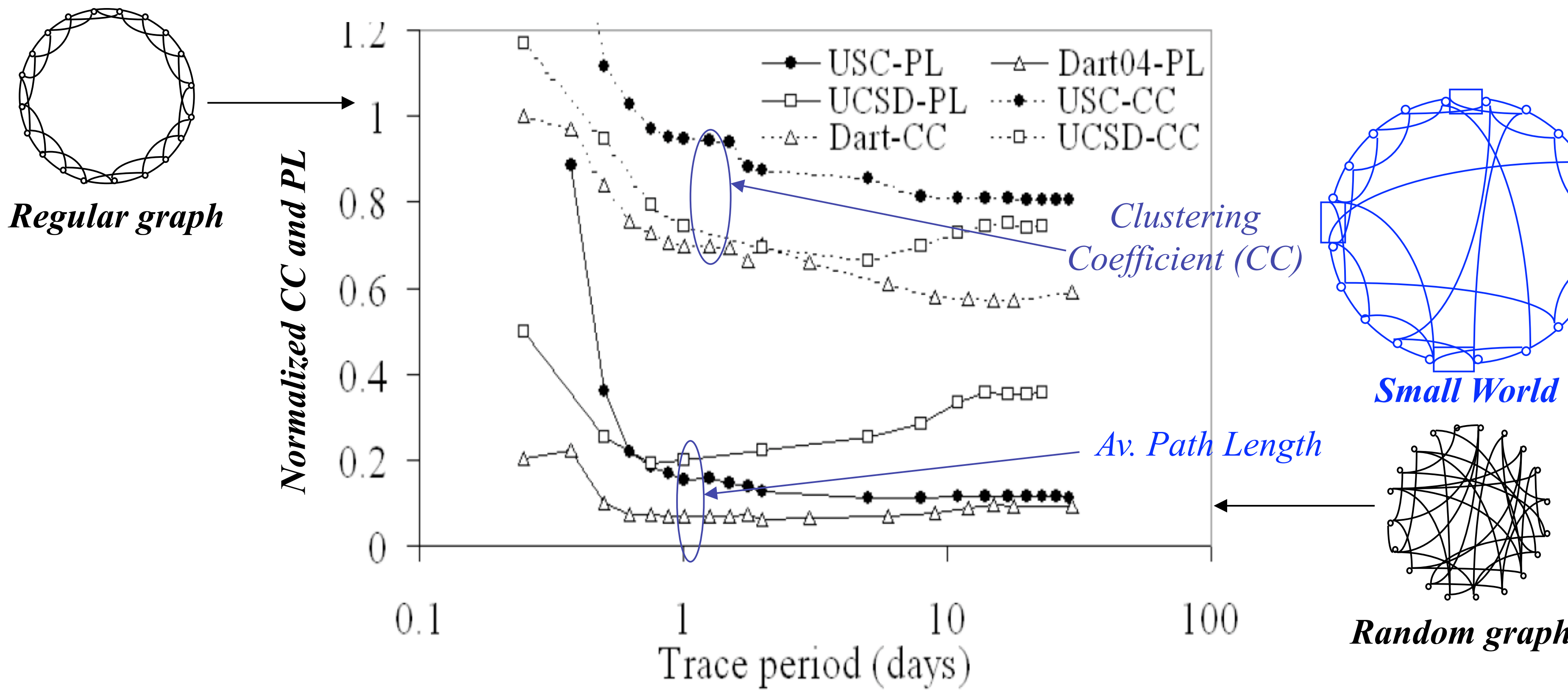


[H03]

- In *Small Worlds*, a few *short cuts* contract the diameter (i.e., path length) of a regular graph to resemble diameter of a random graph without affecting the graph structure (i.e., clustering)

Small Worlds of Encounters

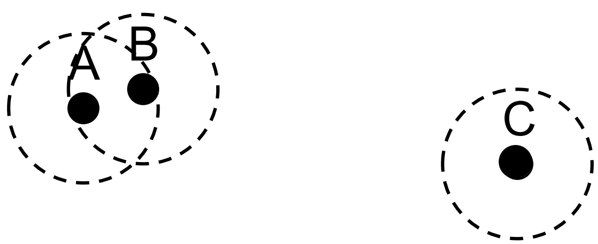
- **Encounter graph: nodes as vertices and edges link all vertices that encounter**



- **The encounter graph is a *Small World* graph (high *CC*, low *PL*)**
- **Even for short time period (1 day) its metrics (*CC*, *PL*) almost saturate**

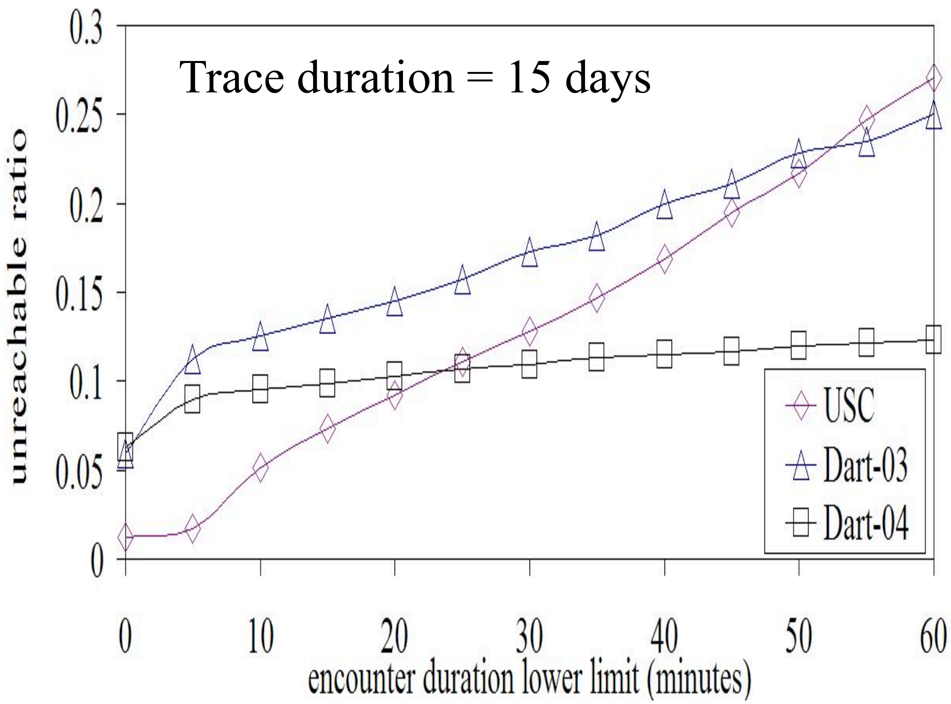
Background: Epidemic Routing in Delay Tolerant Networks (*DTN*)

- *DTNs* are mobile networks with sparse, intermittent nodal connectivity
- Encounter events provide the communication opportunities among nodes
- Messages are stored and moved across the network with nodal mobility

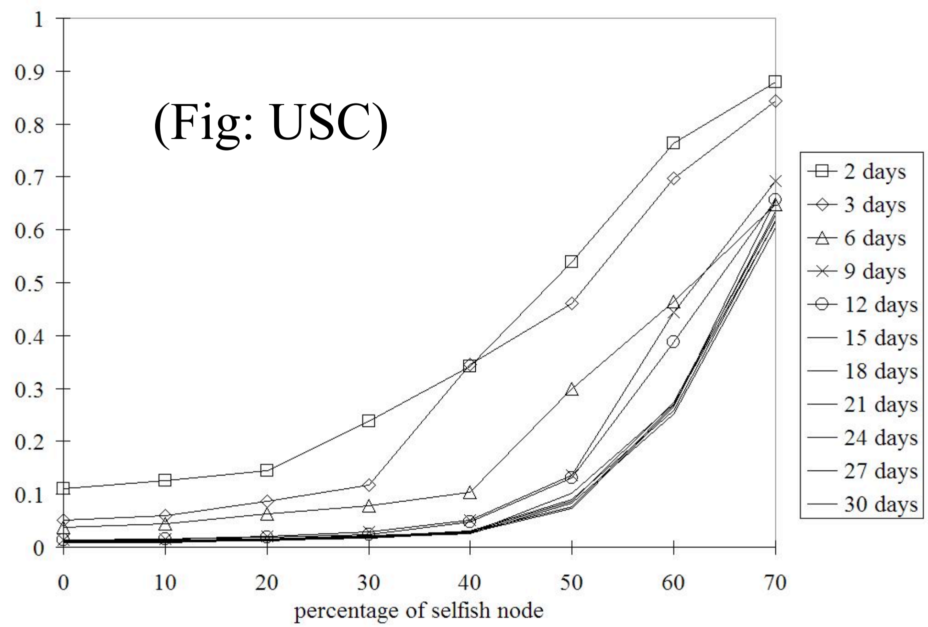


Information Diffusion in DTNs via Encounters

- Epidemic routing (spatio-temporal broadcast) achieves almost complete delivery



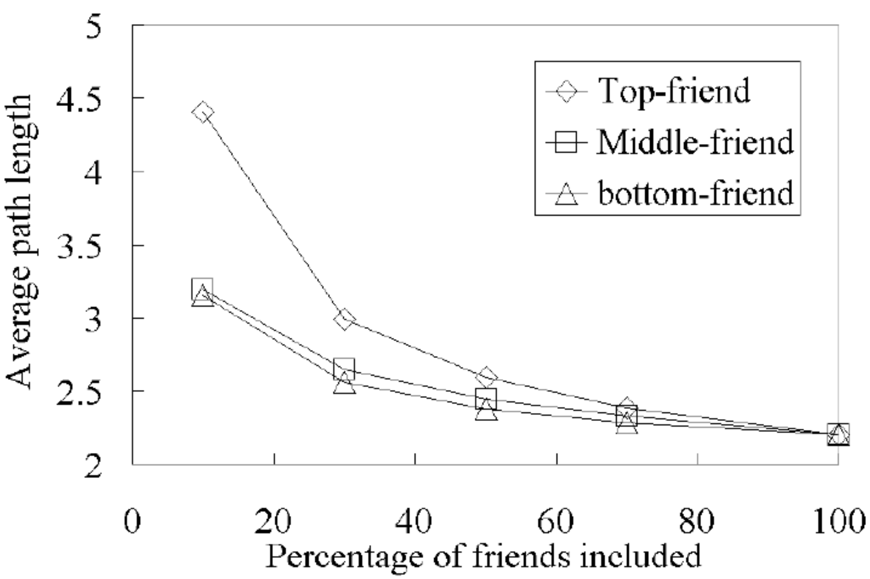
Robust to the removal of short encounters



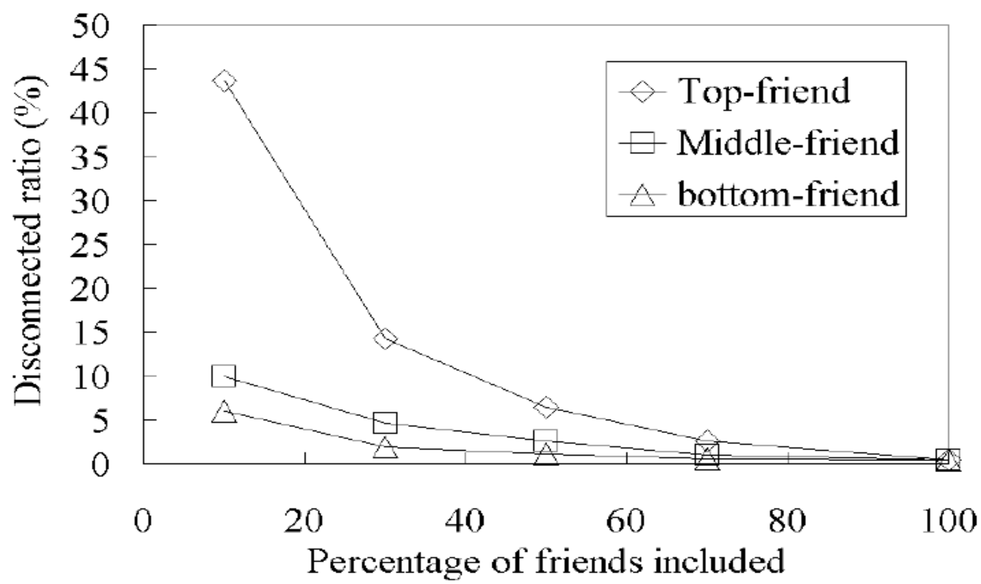
Robust to selfish nodes (up to ~40%)

Encounter-graphs using Friends

- **Distribution for friendship index *FI* is exponential for all the traces**
- **Friendship between MNs is highly asymmetric**
- **Among all node pairs: < 5% with *FI* > 0.01, and <1% with *FI* > 0.4**



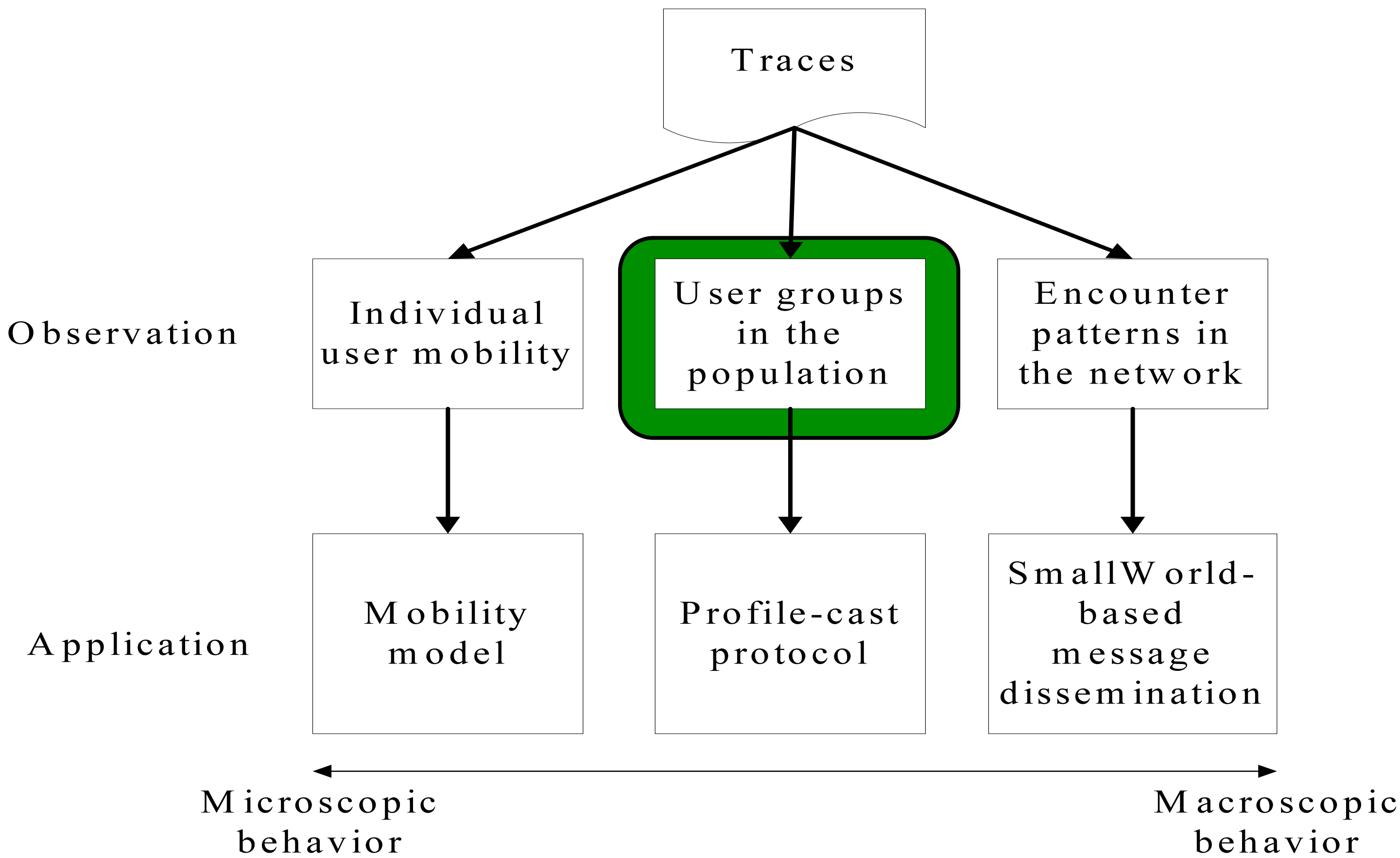
(b) Average path length



(c) Disconnected ratio

•Top-ranked friends form cliques and low-ranked friends are key to provide random links (short cuts) to reduce the degree of separation in encounter graph.

Case study III – Groups in WLAN



Case Study III: Goal

- Identify similar users (in terms of long run mobility preferences) from the diverse WLAN user population
  - Understand the constituents of the population
  - Identify potential groups for group-aware service
- Classify users based on their mobility trends and location-visiting preferences
  - Traces studied: semester-long USC trace (spring 2006, 94days) and quarter-long Dartmouth trace (spring 2004, 61 days)

Representation of User Association Patterns

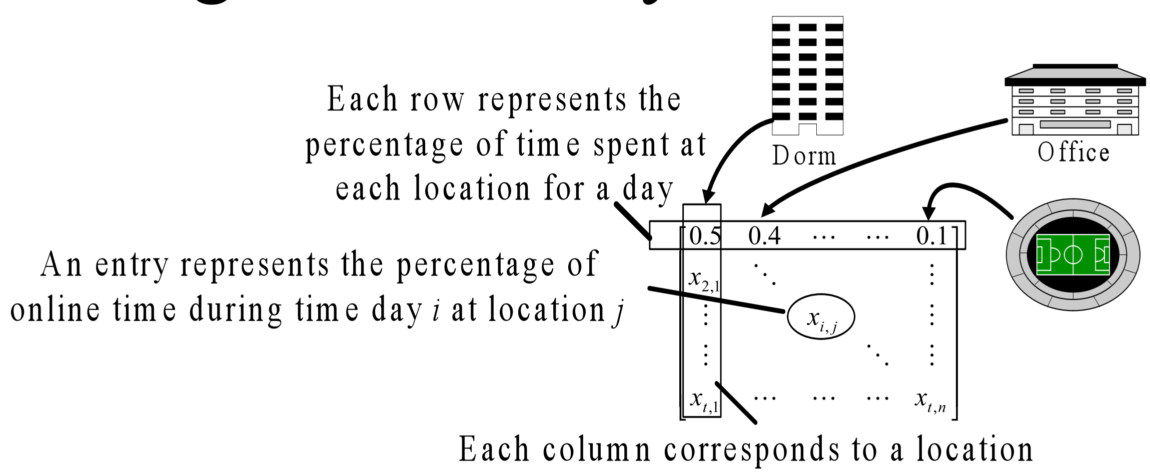
W. Hsu, D. Dutta, A. Helmy, “Mining Behavioral Groups in WLANs”, *ACM MobiCom* ‘07

- Summarize user association per day by a vector
  - $a = \{a_j$: fraction of online time user i spends at AP_j on day $d\}$

-Office, 10AM -12PM
-Library, 3PM – 4PM
-Class, 6PM – 8PM

Association vector:
(library, office, class) =(0.2, 0.4, 0.4)

- Sum long-run mobility in “association matrix”



Example association matrix to describe a given user’s location visiting preference

$$\begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{t,1} & \cdots & x_{t,n} \end{bmatrix}$$

Represent

Association Matrix

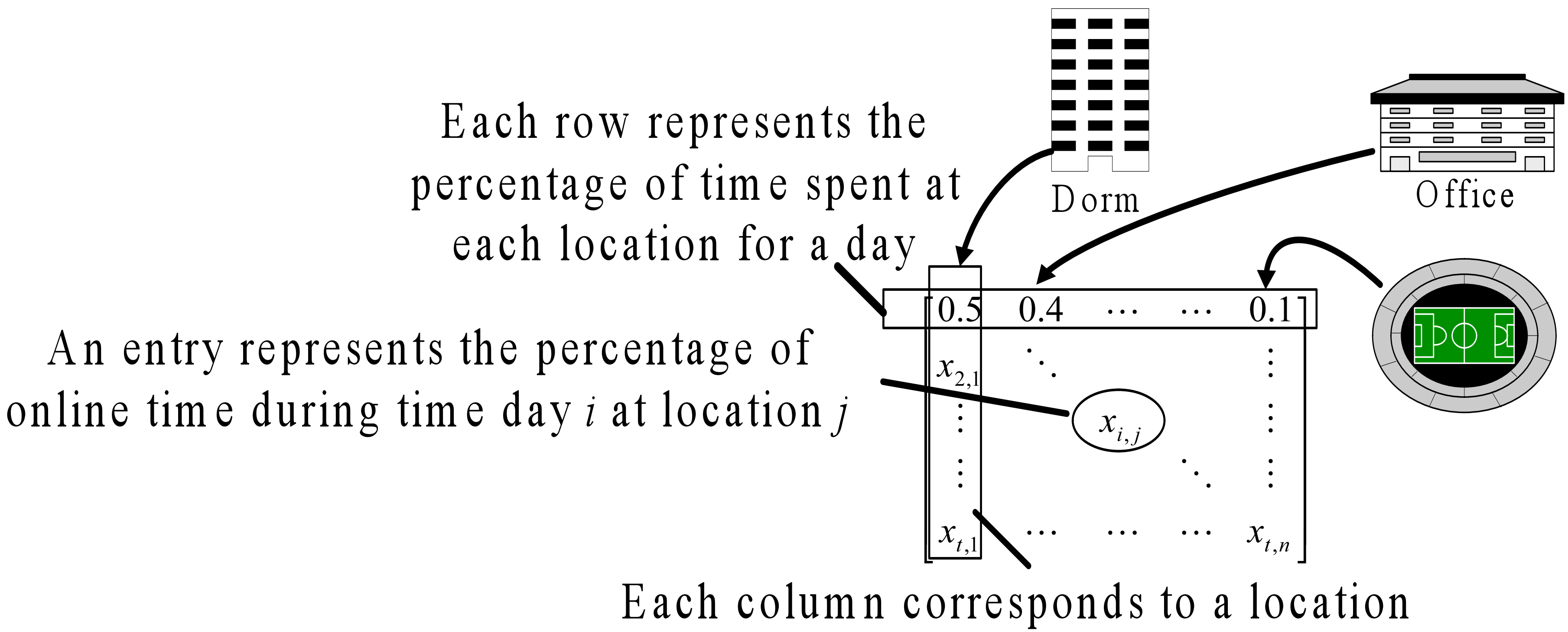


Illustration of the association matrix to describe a given user's location visiting preference.

Eigen-behaviors & Behavioral Similarity Distance

- Eigen-behaviors (*EB*): Vectors describing maximum remaining power in assoc. matrix *M* (through *SVD*):

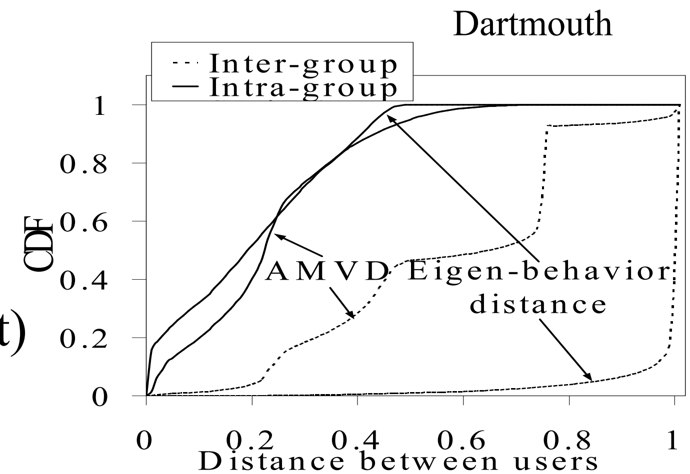
$$M = U \cdot \Sigma \cdot V^T$$

  - Get Eigen-vectors: $v_1, v_2, ..., v_{rank(M)}$ - Get Eigen-values: $\sigma_1,\ \sigma_2,\ ...,\ \sigma_{rank(M)}$
  - Get relative importance: $w_i = \frac{\sigma_i^2}{\sum_{j=1}^{Rank(M)} \sigma_j^2}$

- Eigen-behavior Distance weighted inner products of *EBs*
  - $Sim(U,V) = \sum_{\forall i,j} w_i w_j \left| u_i \cdot v_j \right|$
- Assoc. patterns can be re-constructed with low rank & error
- For over 99% of users, < 7 vectors capture > 90% of *M*'s power

Similarity-based User Classification

- Hierarchical clustering of similar behavioral groups
- High quality clustering:
  - Inter-group vs. intra-group distance
  - Significance vs. random groups
    - 0.93 v.s. 0.46 (USC), 0.91 v.s. 0.42 (Dart)



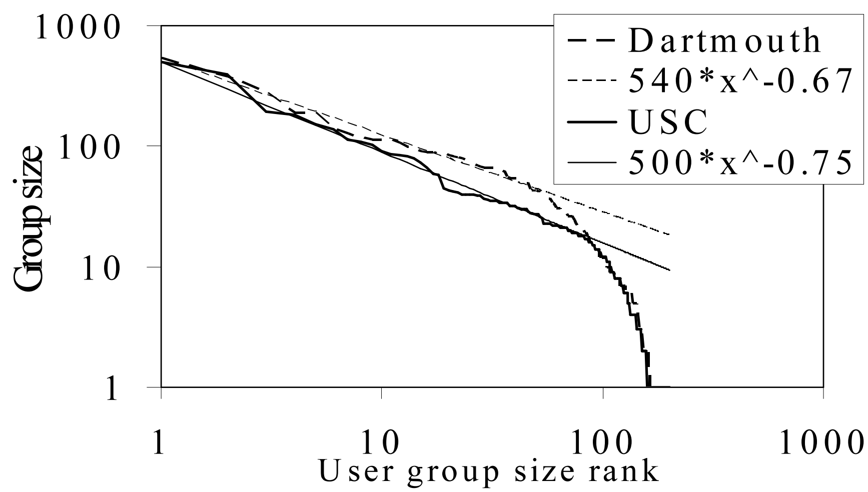
*AMVD = Average Minimum Vector Distance

  - Unique groups based on *Eigen Behaviors*

Significance score of top eigen-behavior for	USC	Dartmouth
Its own group	0.779	0.727
Other groups	0.005	0.004

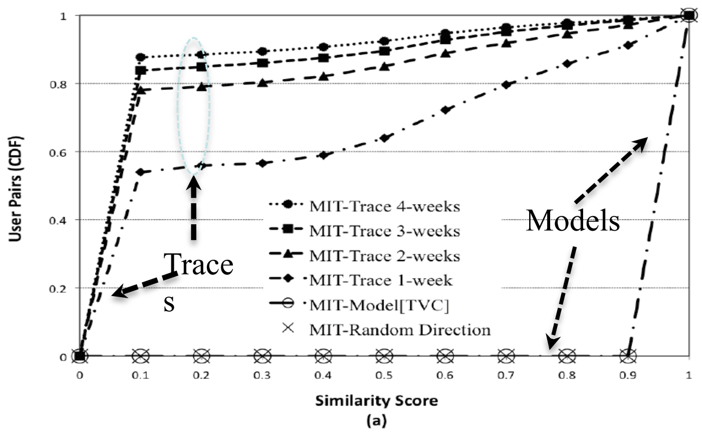
User Groups in WLAN - Observations

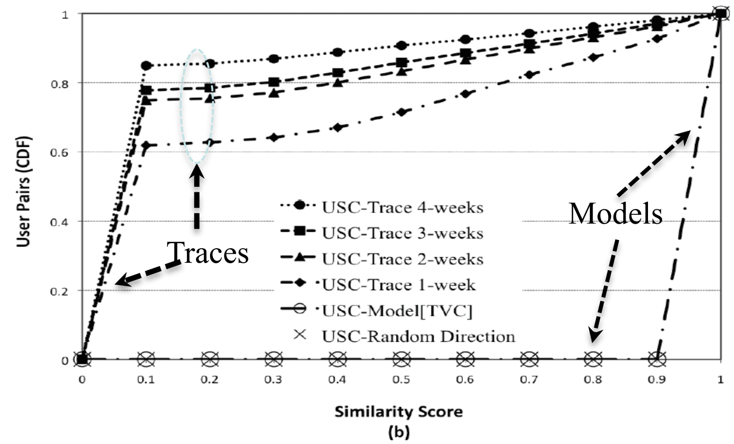
- Identified hundreds of distinct groups of similar users
- Skewed group size distribution –
  - the largest 10 groups account for more than 30% of population on campus
  - *Power-law distributed of cluster group sizes*
- Most groups can be described by a list of locations with a clear ordering of importance
- Some groups visit multiple locations with similar importance –
  - taking the most important location for each user is not sufficient



Videos

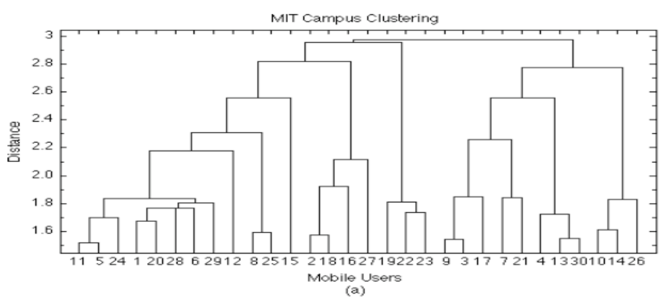
Behavioral Similarity: The Missing Link

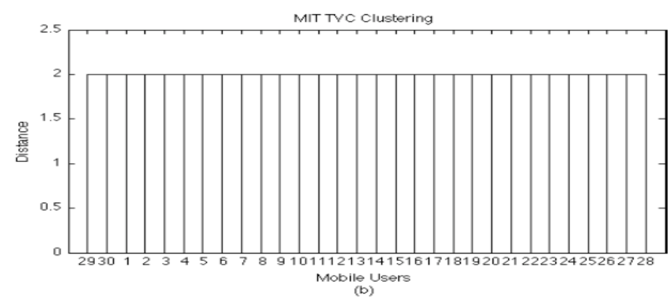


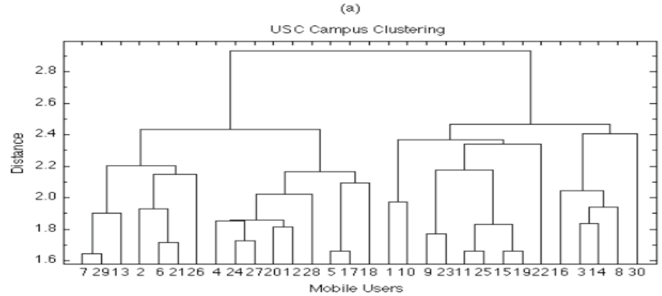


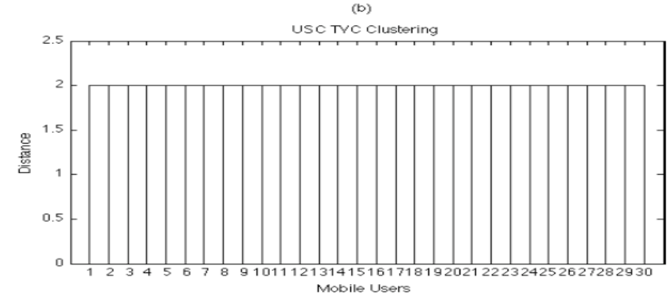
Traces

Models



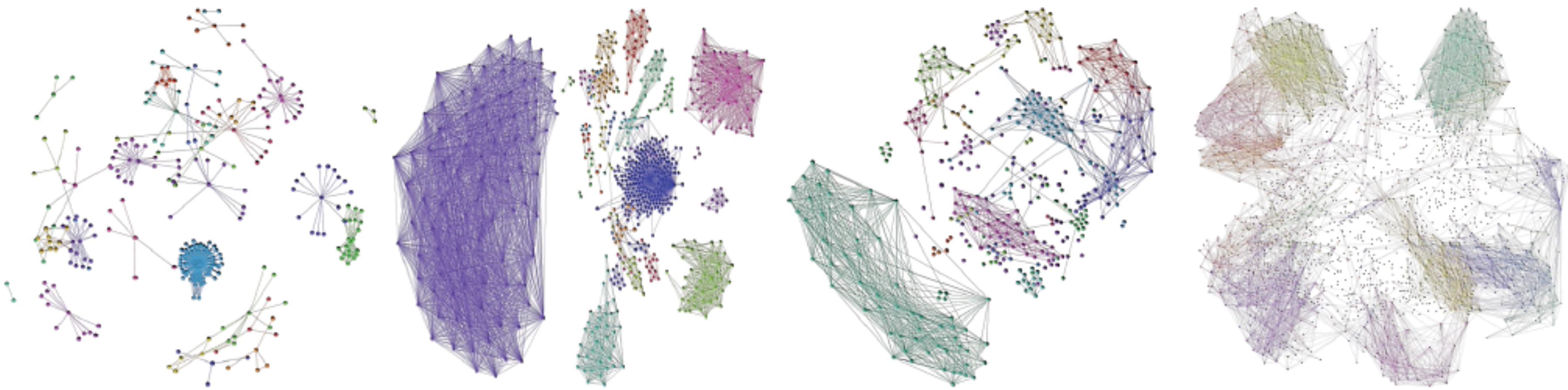




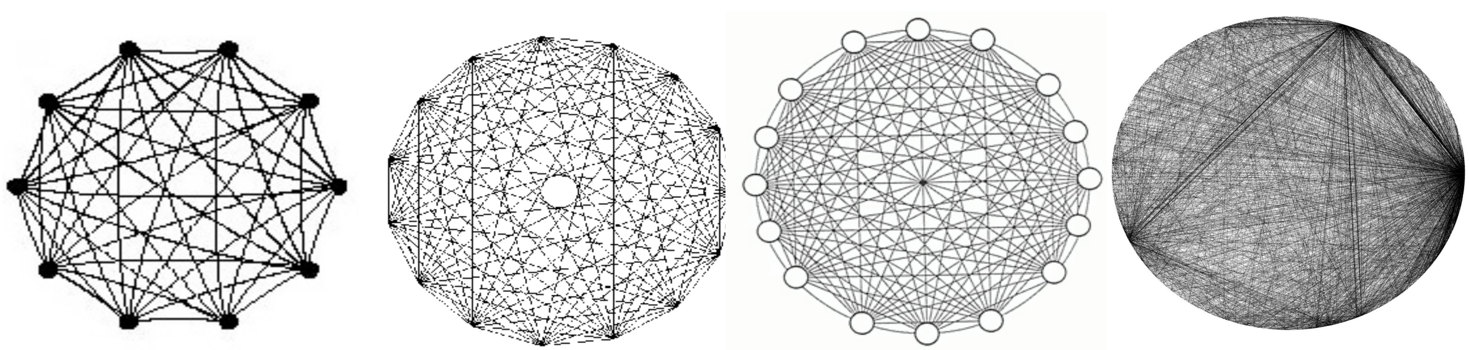


Existing models produce behaviorally homogeneous users and lack the richness of behavioral structure in real traces. Richer models are needed !

Behavioral Similarity Graphs



(a) Dartmouth Campus (b) MIT Campus (c) UF Campus (d) USC Campus



Random and community models produce fully connected similarity graphs

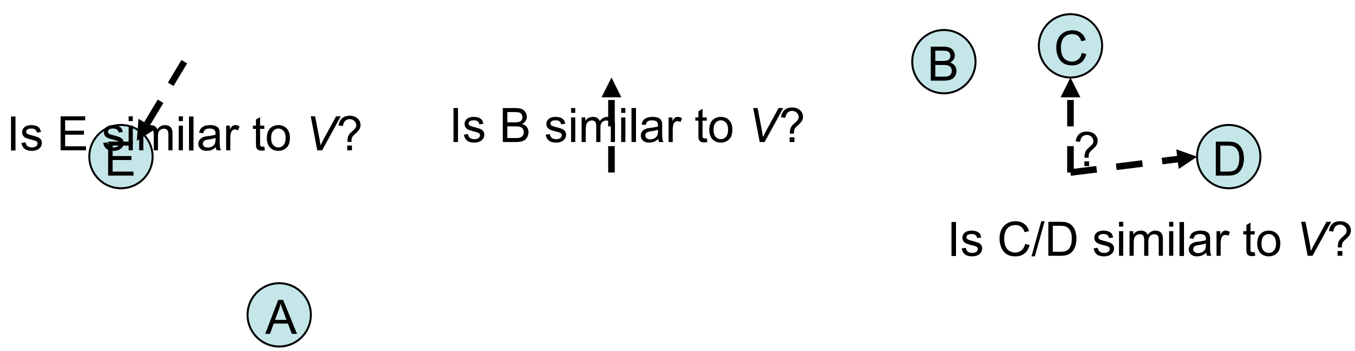
G. Thakur, A. Helmy, W. Hsu, “Similarity analysis and modeling of similarity in mobile societies: The missing link”, *ACM CHANTS*, Sep 2010

Profile-cast: A New Communication Paradigm

W. Hsu, D. Dutta, A. Helmy, *ACM Mobicom* 2007, *WCNC* 2008, *Journals under submission*

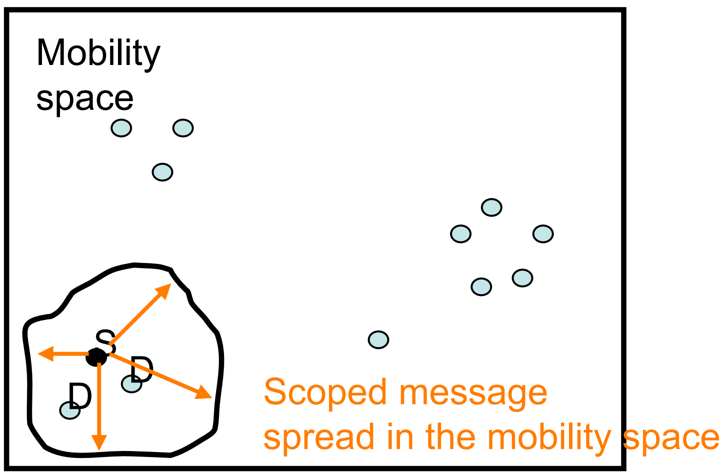
Payload	Dest Address	→	Payload	Target Profile

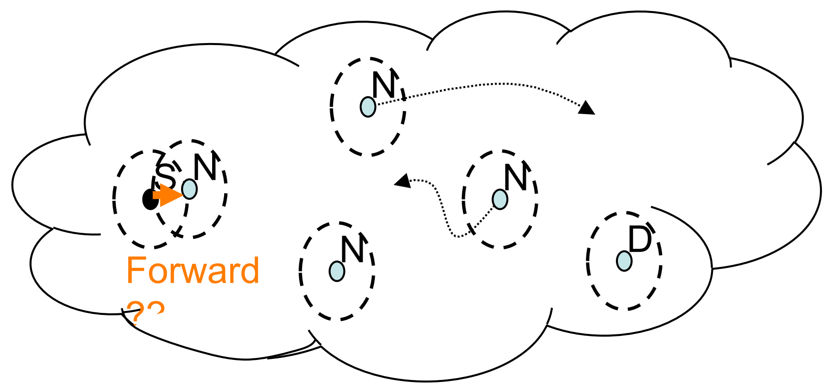
- Sending messages to others with similar behavior, without knowing their identity
  - Announcements to users with specific behavioral profile *V*
  - Interest-based ads, similarity resource discovery
  - A novel paradigm for "*Behavior-based Networking*"
- Example application in Delay Tolerant Networks (DTNs)



Profile-cast Use Cases

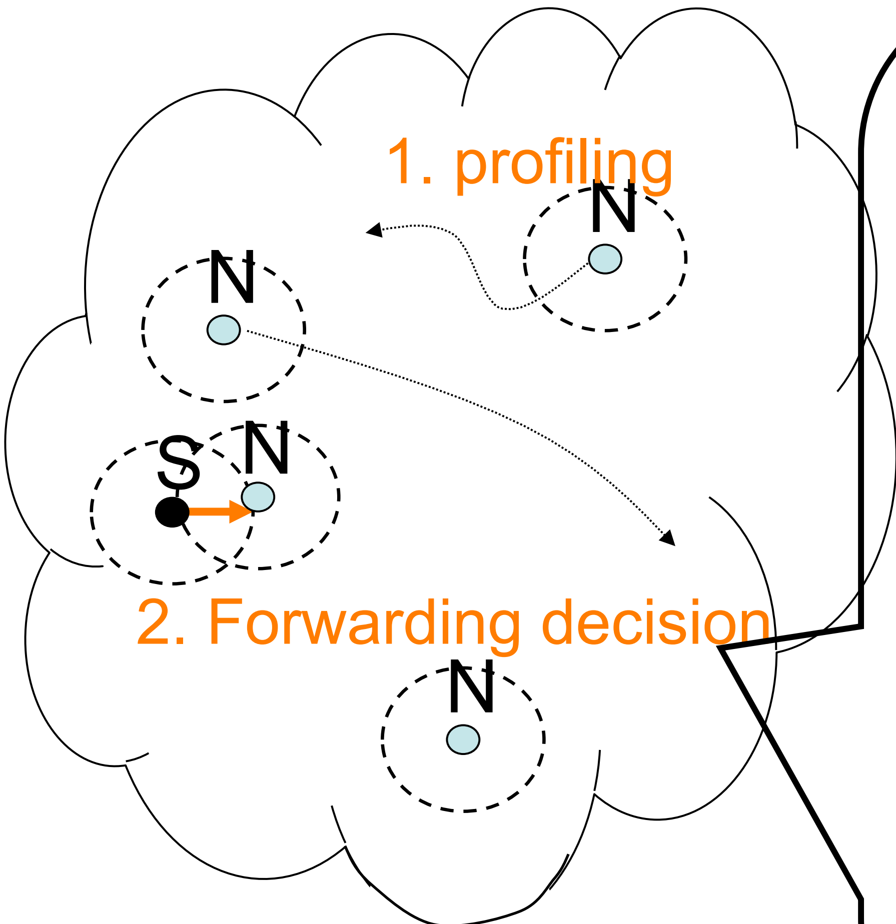
- Mobility-based *profile-cast* (Target mode)
  - Targeting group of users who move in a particular pattern (lost-and-found, context-aware messages, moviegoers)
  - Approach: use "similarity metric" between users





- Mobility-independent *profile-cast* (Dissemination mode)
  - Targeting people with a certain characteristics independent of mobility (classic music lovers)
  - Approach: use "Small World" encounter patterns

Profile-cast Operation

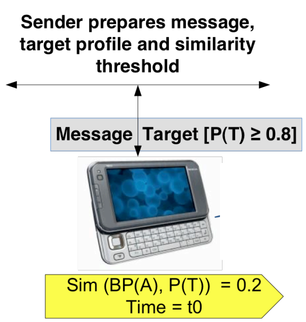


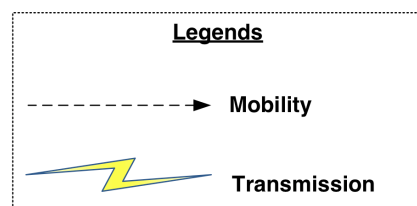
- Determining user similarity
  - ***S*** sends Eigen behaviors for the *virtual* profile to ***N***
  - ***N*** evaluated the similarity by weighted inner products of Eigen-behaviors

$$Sim(U,V) = \sum_{\forall i,j} w_i w_j \left| u_i \cdot v_j \right|$$

  - Message forwarded if Sim(U,V) is high (the goal is to deliver messages to nodes with similar profile)
  - Privacy conserving: ***N*** and ***S*** do not send information about their own behavior

Profile-cast CSI protocol: Target-mode

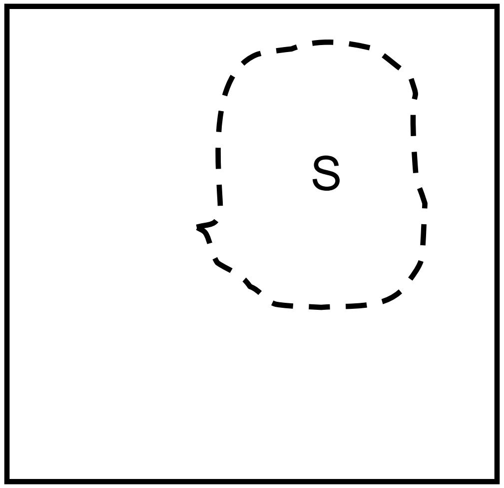




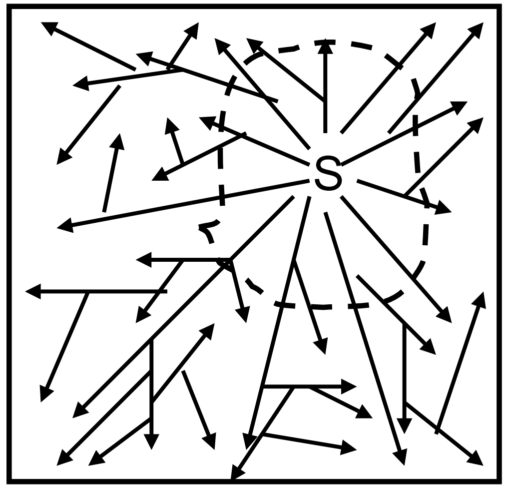
Sim (BP(A), P(T)) = similarity of node’s behavioral profile to the target profile

Mobility Profile-cast (intra-group)

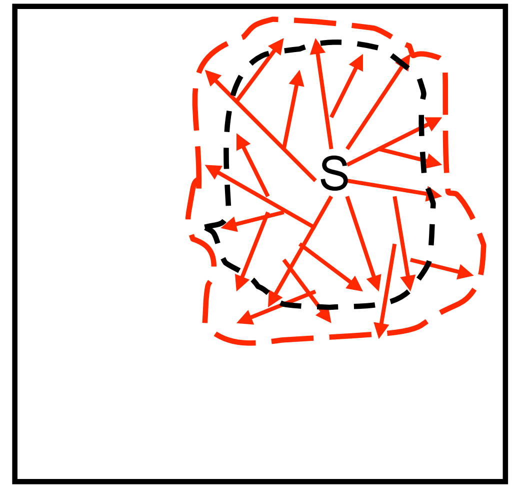
Goal



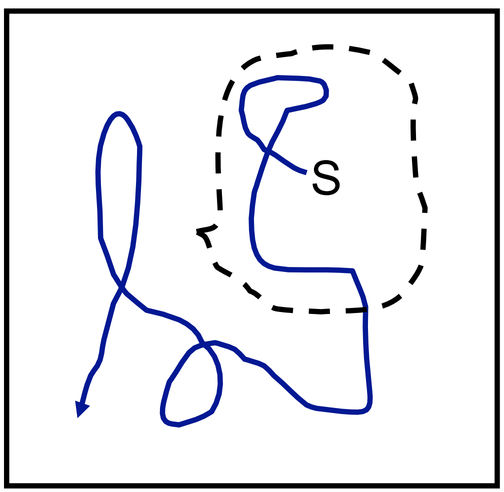
Epidemic



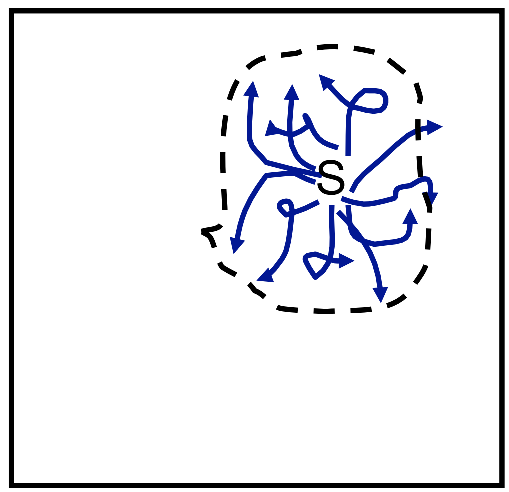
Group-spread



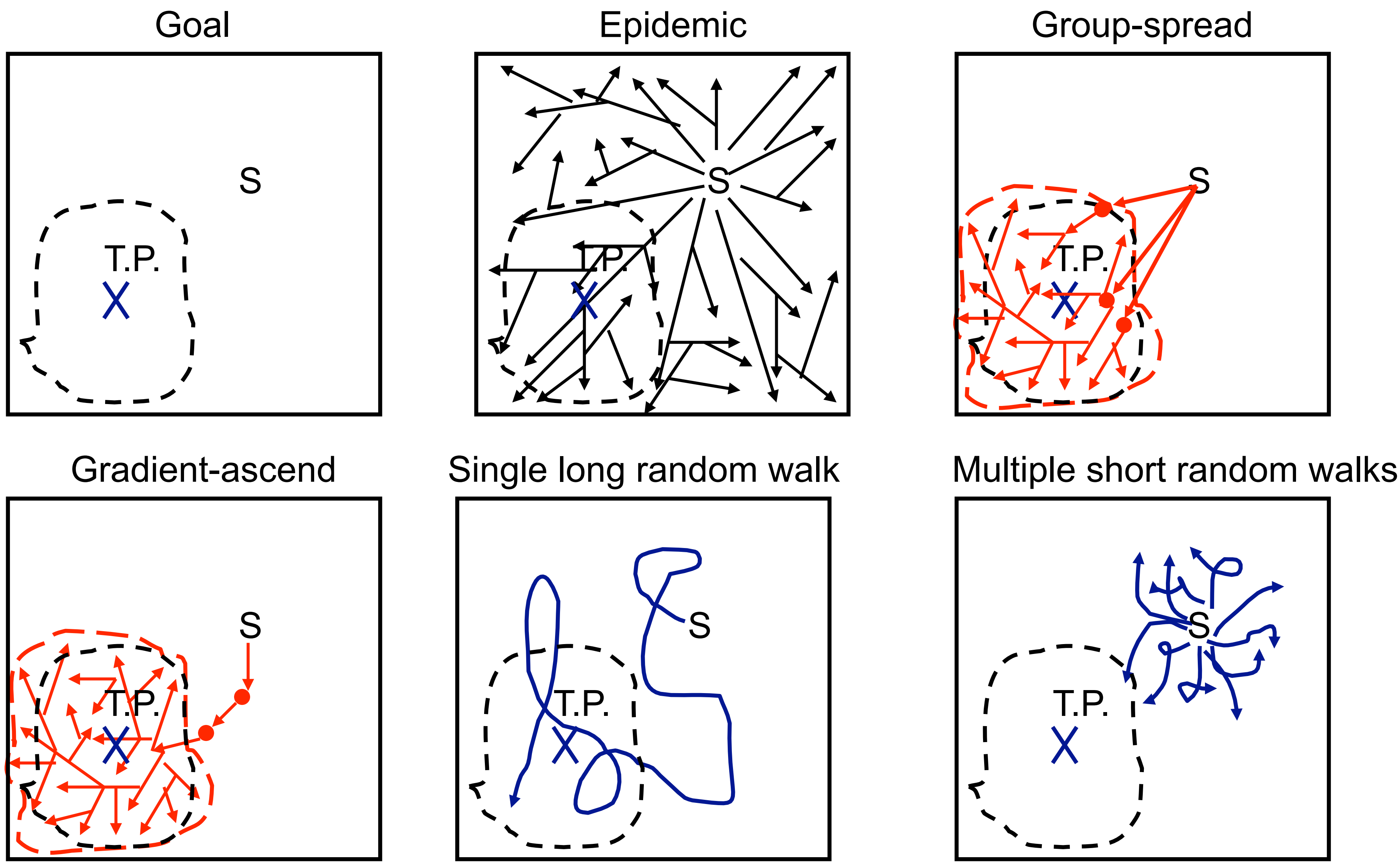
Single long random walk



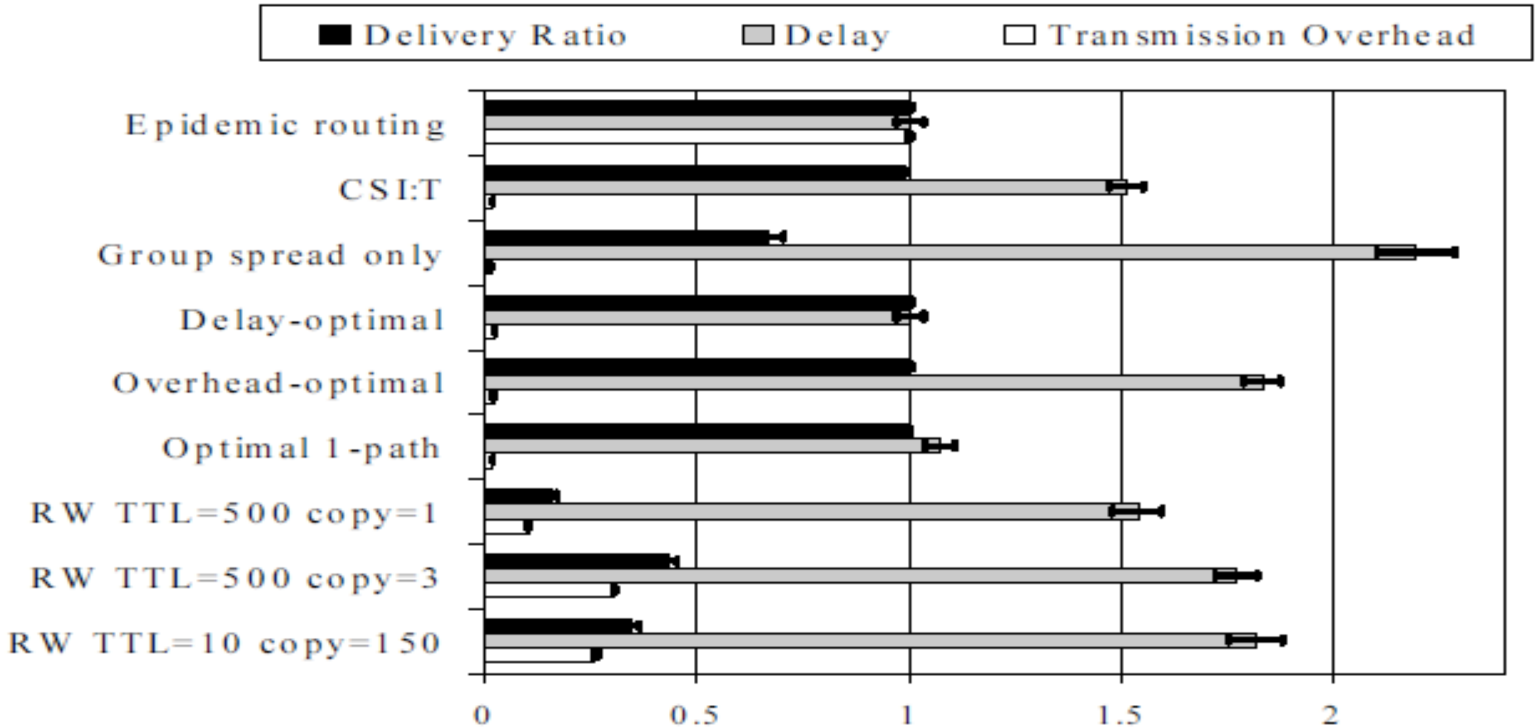
Multiple short random walks



Mobility Profile-cast (inter-group)



Profile-cast Evaluation



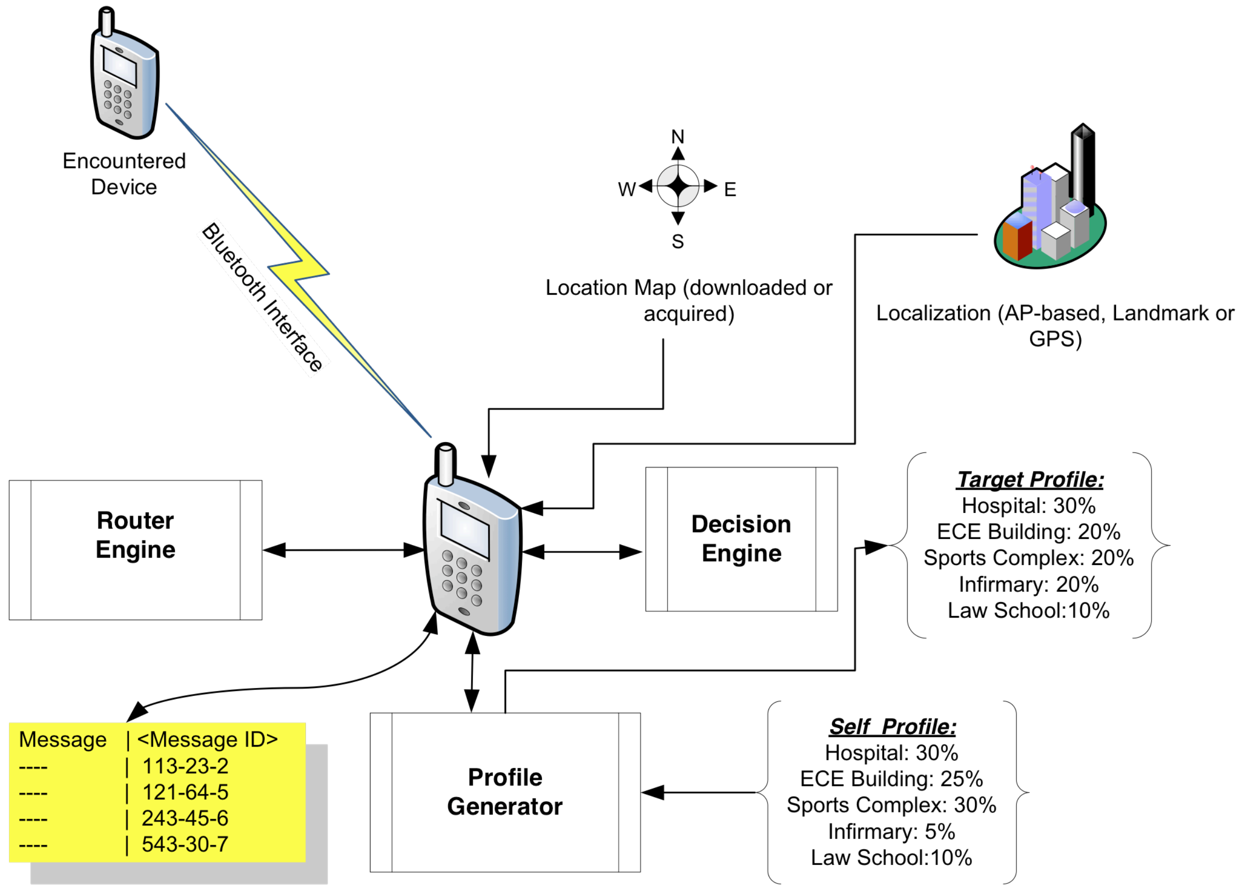
(a) USC.

* Results presented as the ratio to epidemic routing

- Over 96% delivery ratio – Over 98% reduction in overhead w.r.t. Epidemic
- RW < 45% delivery
- Strikes a near optimal balance between delivery, overhead and delay
- Other variants (e.g., multi-copy, simulated annealing) under investigation

Video

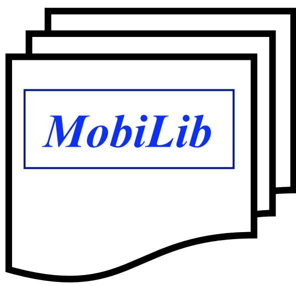
Implementation Details



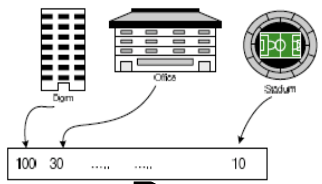
videos

The *TRACE* framework

	Duration	Records	Total Users	Access points (or bldgs)
USC-WLAN	Dec 03-Jun 08	50 M	55,500	79 ports (03), 161 (08)
USC-DHCP	Dec 03-Jun 08	60 M	55,500	79 ports (03), 161 (08)
USC-netflow	Apr 05-Jun 08	50 B	50,000	161 ports
UF-WLAN	Jun 07-Current	45 M	105,500	784 Access points
UF-DHCP	Jun 07-Current	10 M	105,500	784 Access points
UF-netflow	to start Sep 09	n/a	n/a	784+ Access points

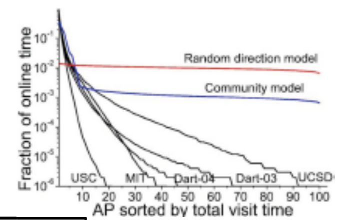


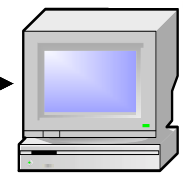
Trace



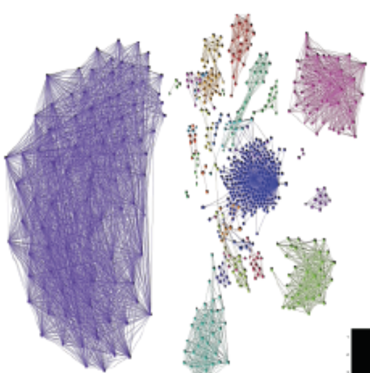
$$\begin{pmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{t,1} & \cdots & x_{t,n} \end{pmatrix}$$

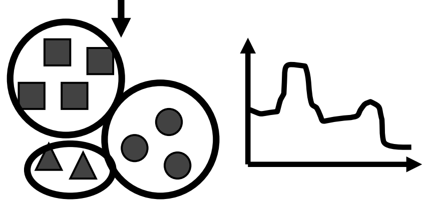
Represent



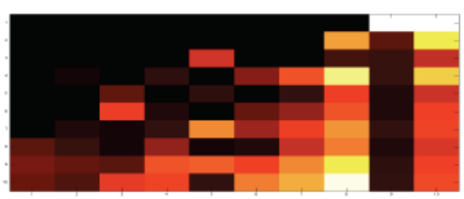


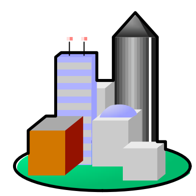
Analyze

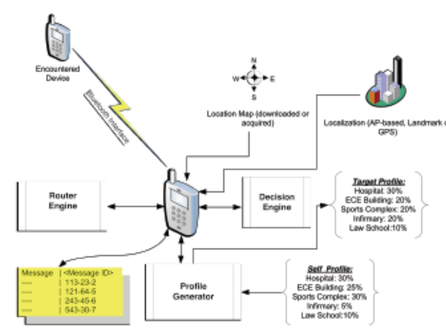




Characterize, **C**luster

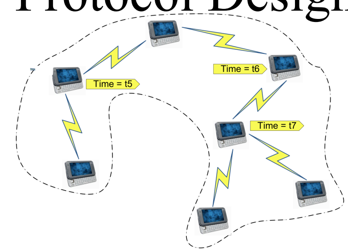






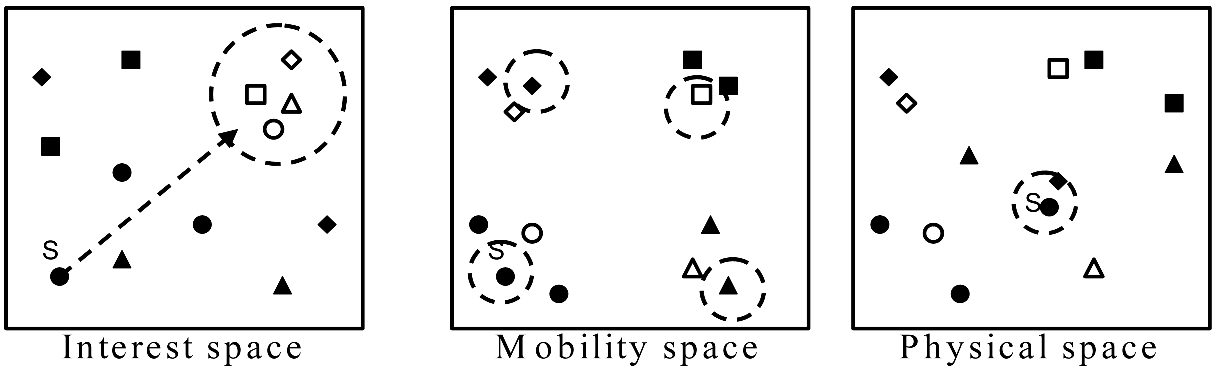
Employ

(Modeling, Protocol Design)



Extensions to *Profile-Cast*: Interest-Cast (*iCast*)

– *Disseminate mode: N*-copy-per-clique in the "mobility space"



- Different legends represent nodes with different mobility trends
- White nodes denote the target recipients

– Challenge: From mobility to interest and other classifications

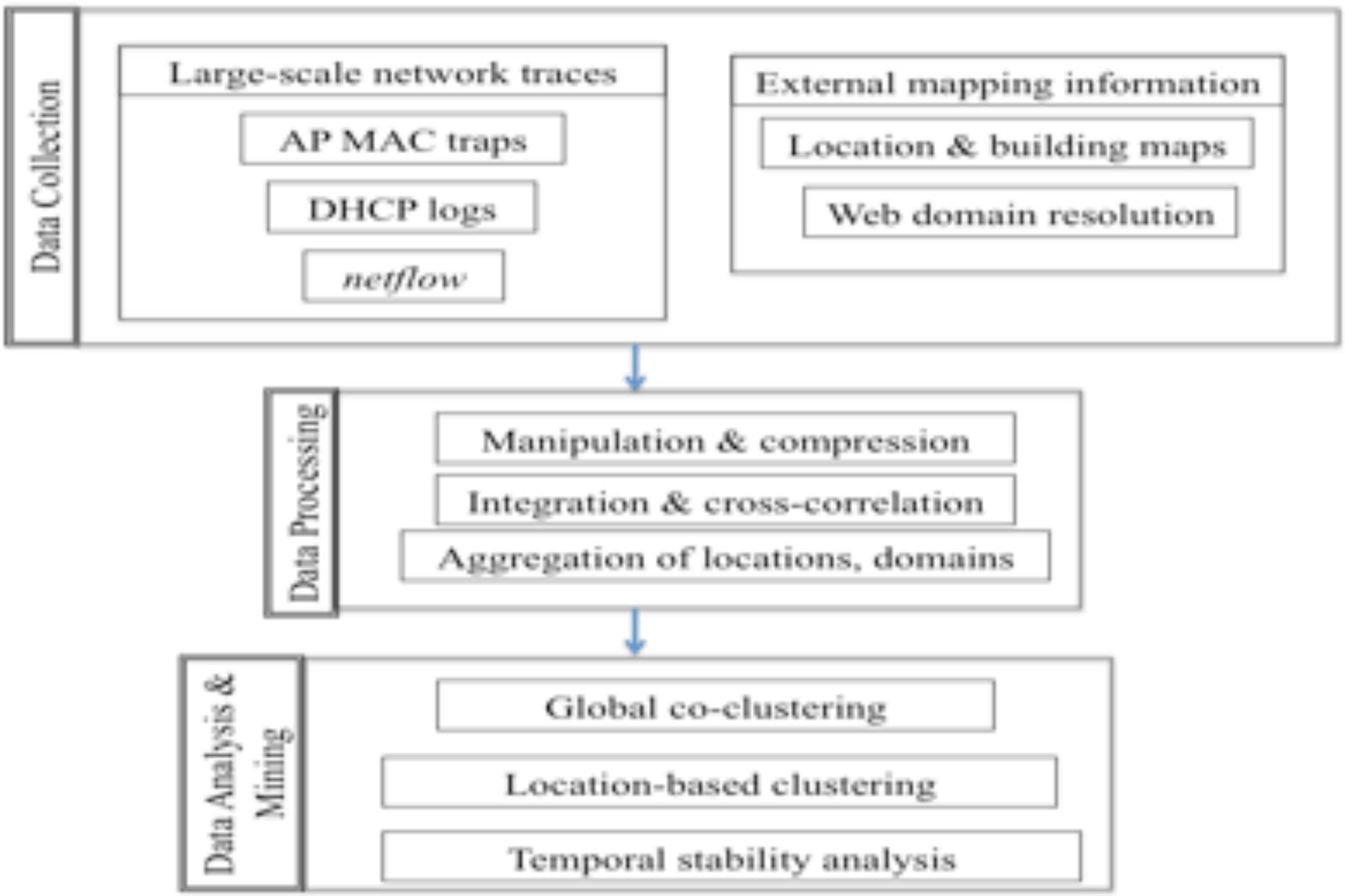
Extending Interest: *Behavior Beyond Mobility*

- *Dimensionality*: In addition to mobility, user's web access and traffic patterns, applications used (among others) represent other (more personal/social) dimensions of interest and behavior
- Further analysis of network measurements (e.g., *Netflow*) can reveal behavioral characteristics in these dimensions
- *Scale*: Netflow traces are 3 orders of magnitude larger than WLANs (*WLANs*: dozens of millions, *Netflows*: dozens of billions)
- New challenges in mining 'big data' to get information

	Duration	Records	Total Users	Access points ports
USC-WLAN	Dec 03-Jun 08	50 M	55,500	79 ports (03), 161 (08)
USC-DHCP	Dec 03-Jun 08	60 M	55,500	79 ports (03), 161 (08)
USC-netflow	Apr 05-Jun 08	50 B	50,000	161 ports
UF-WLAN	Jun 07-Current	60 M	140,000	784 Access points
UF-DHCP	Jun 07-Current	13 M	140,000	784 Access points
UF-netflow	starting Nov	2.5B/month	45,000	784 Access points

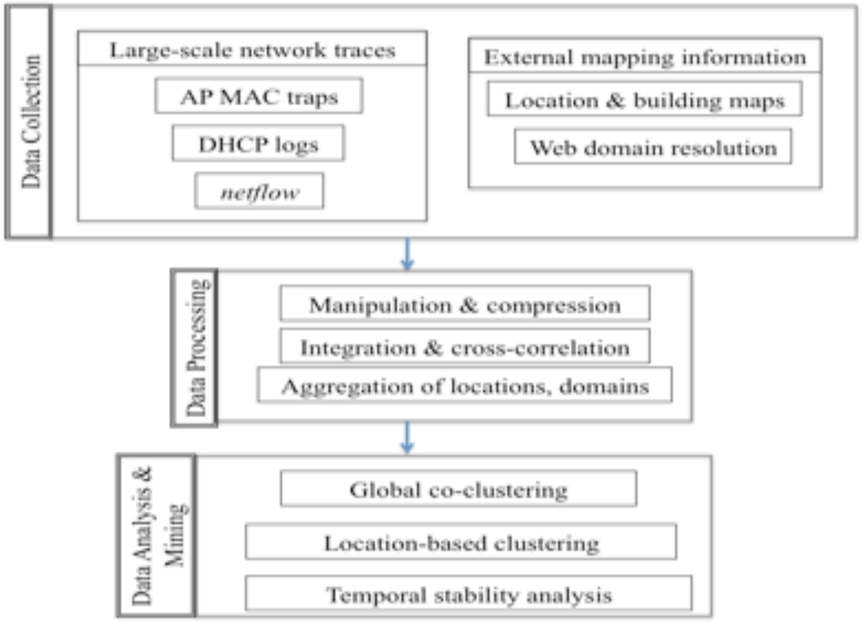
Web Access Analysis Framework

* S. Moghaddam, A. Helmy, S. Ranka, M. Somaya, “Data-driven Co-clustering Model of Internet Usage in Large Mobile Societies”, *ACM MSWIM*, Oct 2010



Data Collection & Processing

- Billions of netflow records
- Cross-correlated with other traces and information
- To find user, accessed domain and building for each record.
- Our case study: over 2 billion flow records for Feb. -Apr. 2008



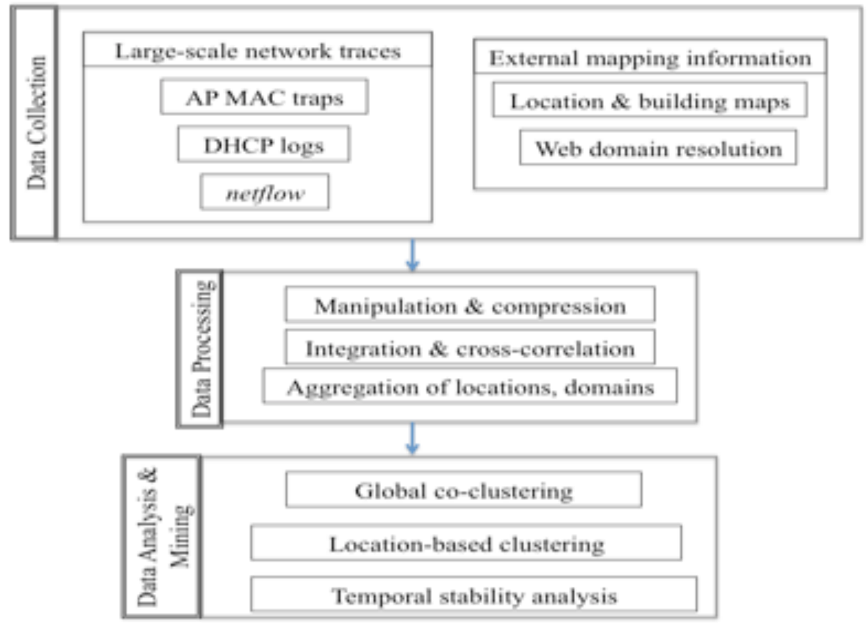
Netflow sample.

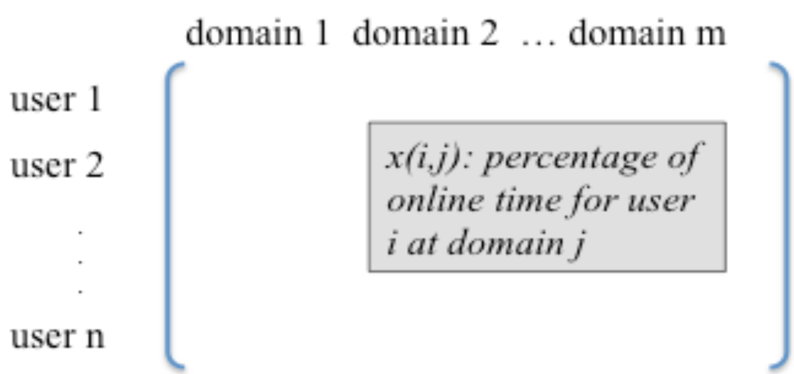
Start Timestamp	Finish Timestamp	Source IP	Source Port	Dest IP	Dest Port	Protocol Num	ToS	Packet Count	Flow Size
0618.00:00:07.184	0618.00:00:07.184	128.125.253.14	53	207.151.25.121	64209	17	0	1	469
0618.00:00:07.184	0618.00:00:07.472	207.151.241.60	52759	74.125.19.17	80	6	0	4	1789
0618.00:00:07.188	0618.00:00:07.188	193.19.82.9	31676	207.151.238.90	43798	17	0	1	103

Data Collection & Processing (contd.)

Aggregation

- Based on the total online time (per minute)
- For each user at different domains.
- Case study: top 100 active domains, 22816 users, 79 buildings.





Co-clustering on users and domains for Mar. 2008

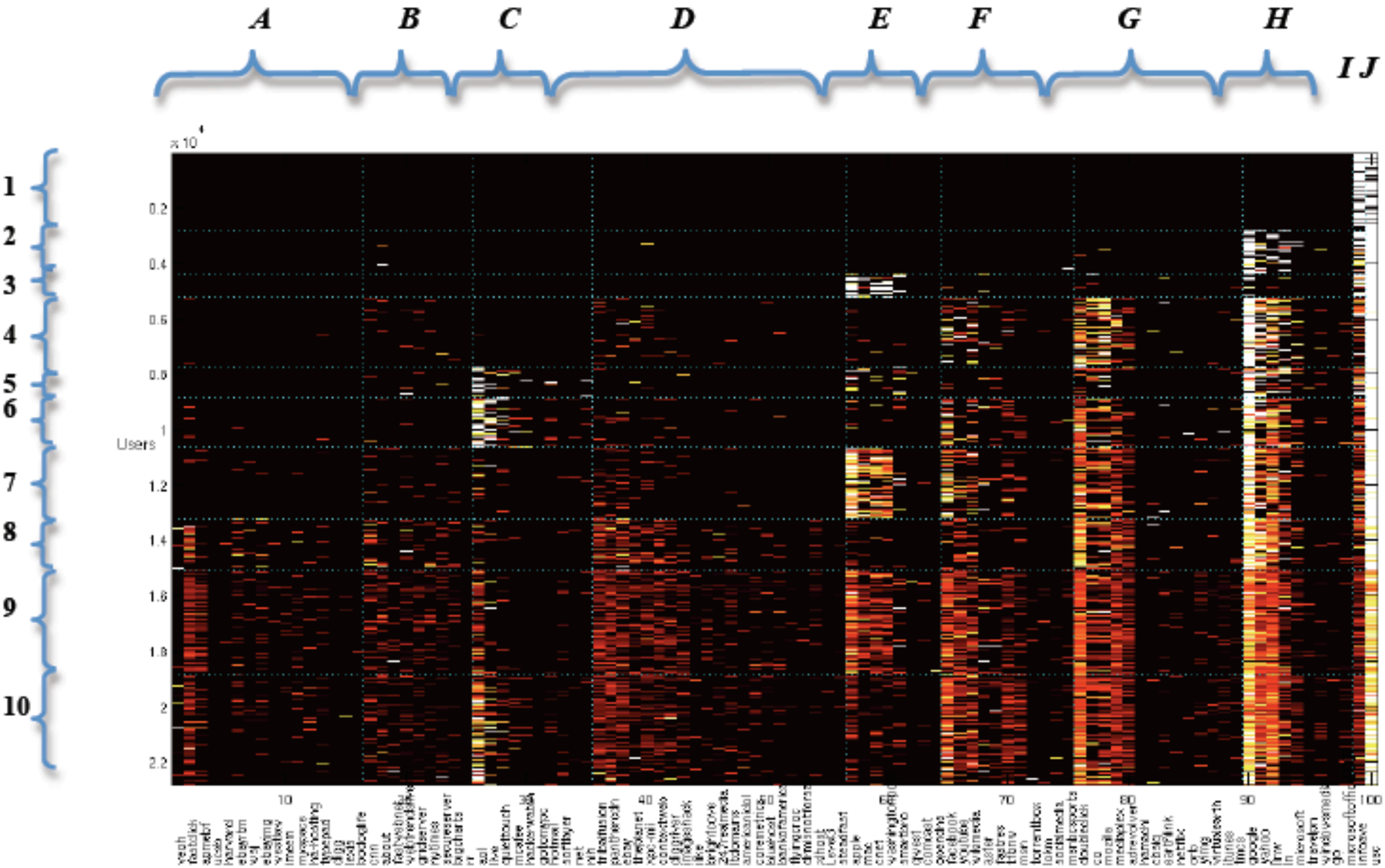
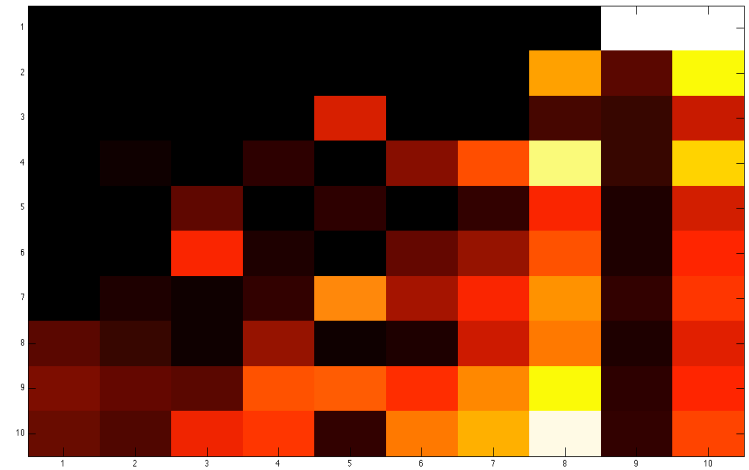


Fig. 5. Information theoretic co-clustering on user-domain matrix, March 2008. The result is given for ten clusters of users (1 through 10) and ten clusters of domains (*A* through *J*). Domain clusters *I* and *J* include one domain each.

Global Data Modeling & Analysis

Major clustered website domains

Cluster	**Domains**
A	myspace – imeem (social media service) - digg (social news) – typepad (blogging) - ebayrtm - ebayimg - wsj (business news) -bodoglife (online gambling) - ucsb - harward - westlaw
B	cnn – nytimes (new york times)
C	mcafee – hackerwatch - live - hotmail
D	ebay - bankofamerica
E	apple – mac - washingtonpost - cnet
F	facebook – youtube - social media msn - msnbcsports
G	netflix – itunes - orb (media cast) - tmcs (social city search) - virtualearth (online map)
H	google – yahoo - microsoft – windowsmedia microsoftoffice2007



Association level matrix: Shows users' behavioral groups based on domain clusters.

- Users can be modeled using few (~10) clusters with clearly disjoint and meaningful profiles. This behavior is very stable (with ~90% similarity) from month to month.

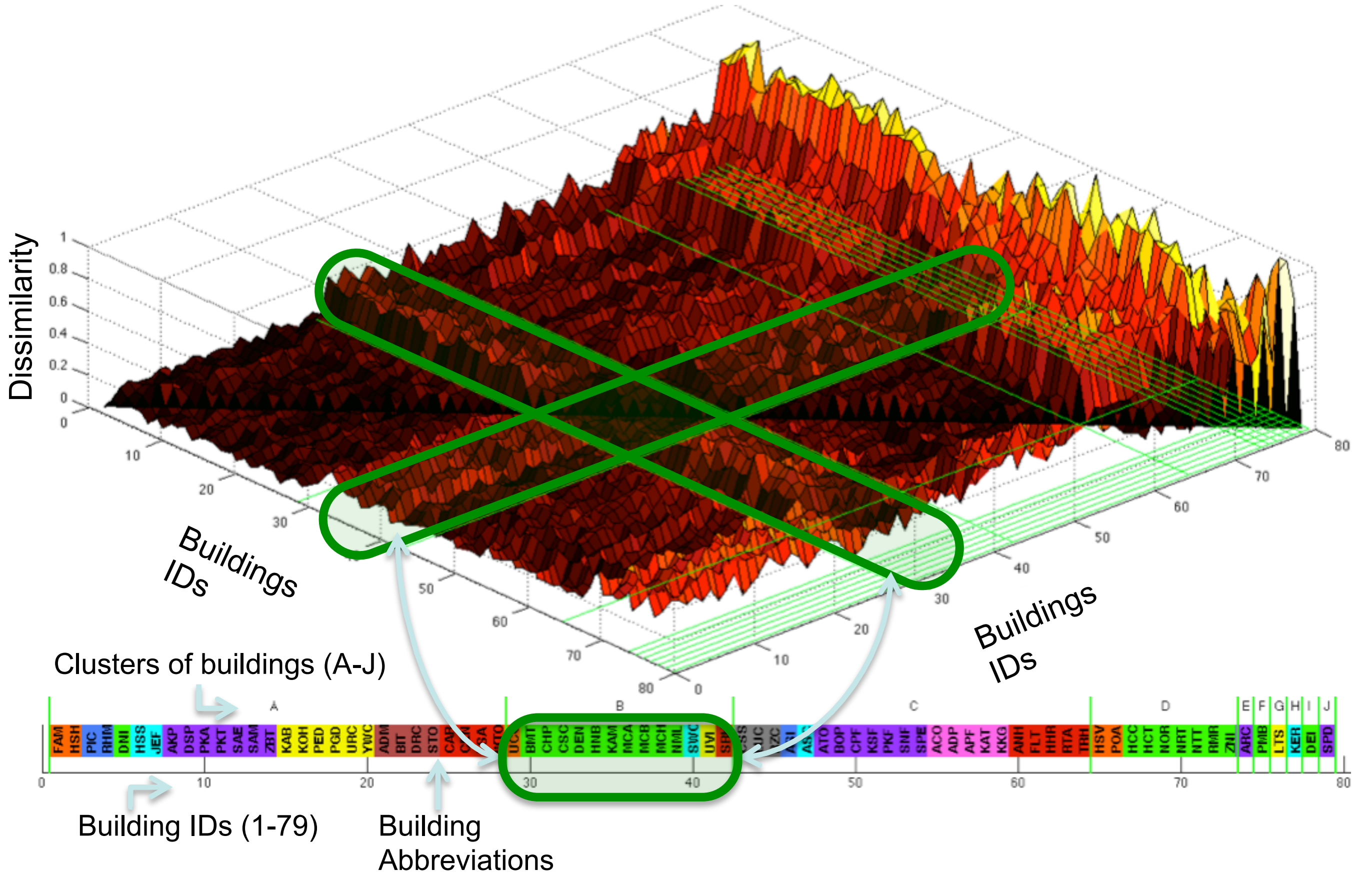
UF UNIVERSITY of FLORIDA

Web-usage Spatio-temporal multi-D Clustering

Location-based Analysis: Hierarchical Clustering of buildings

Building categories based on actual context

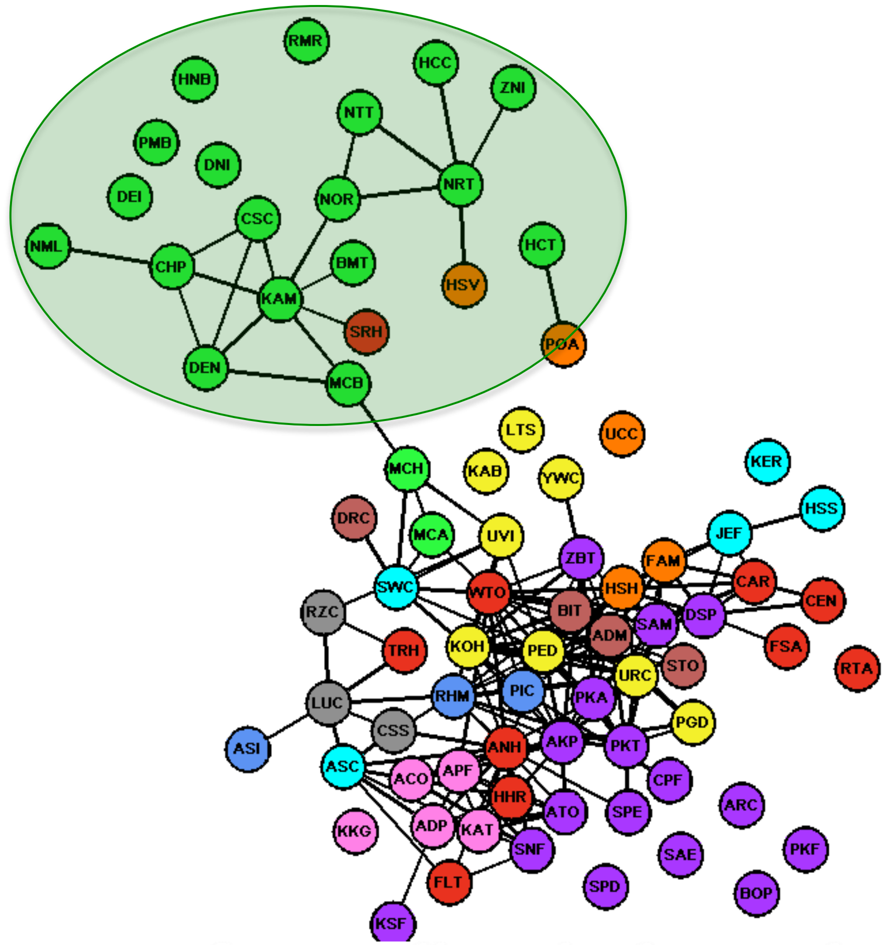
Category	Building Abbreviation				
Activity	KAB	KOH	LTS	PED	PGD
	URC	UVI	YWC		
Auditorium	ADM	BIT	DRC	STO	
Cinema	CSS	LUC	RZC		
Fraternity	AKP	ARC	ATO	BOP	CPF
	DSP	KSF	PKA	PKF	PKT
	SAE	SAM	SNF	SPD	SPE
	ZBT				
Health	BMT	CHP	CSC	DEI	DEN
	DNI	HCC	HCT	HNB	KAM
	MCA	MCB	MCH	NML	NOR
	NRT	NTT	PMB	RMR	ZNI
Housing	ANH	CAR	CEN	FLT	FSA
	HHR	RTA	SRH	TRH	WTO
Music	ASI	PIC	RHM		
School	ASC	HSS	JEF	KER	SWC
Service	FAM	HSH	HSV	POA	UCC
Sorority	ACO	ADP	APF	KAT	KKG



Many of buildings in the same category are clustered together. This trend is table from month to month (Feb through Apr)

Location-based Clique Analysis

Graph representation of dissimilarity matrix for different buildings using the threshold of 0.06.



Category	Building Abbreviation				
Activity	KAB	KOH	LTS	PED	PGD
	URC	UVI	YWC		
Auditorium	ADM	BIT	DRC	STO	
Cinema	CSS	LUC	RZC		
Fraternity	AKP	ARC	ATO	BOP	CPF
	DSP	KSF	PKA	PKF	PKT
	SAE	SAM	SNF	SPD	SPE
	ZBT				
Health	BMT	CHP	CSC	DEI	DEN
	DNI	HCC	HCT	HNB	KAM
	MCA	MCB	MCH	NML	NOR
	NRT	NTT	PMB	RMR	ZNI
Housing	ANH	CAR	CEN	FLT	FSA
	HHR	RTA	SRH	TRH	WTO
Music	ASI	PIC	RHM		
School	ASC	HSS	JEF	KER	SWC
Service	FAM	HSH	HSV	POA	UCC
Sorority	ACO	ADP	APF	KAT	KKG

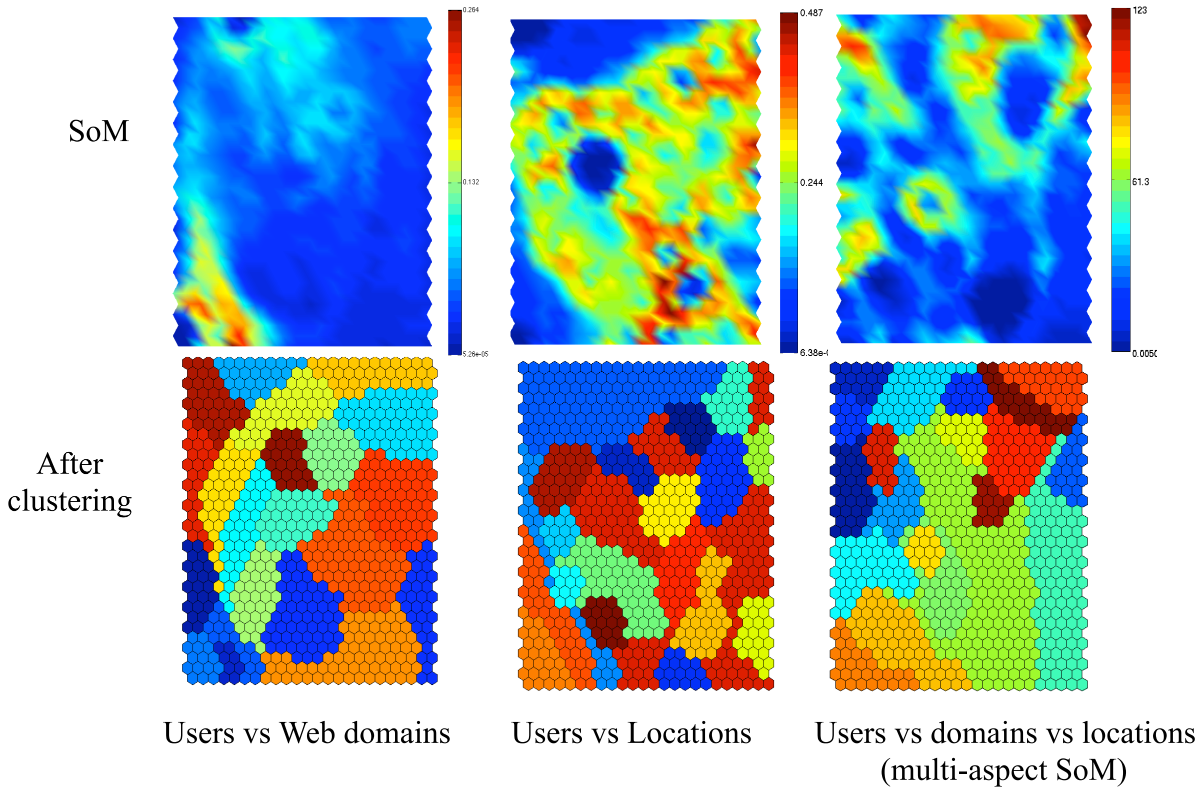
Most buildings in the same category form a clique in the graph

Future Work on *Netflow* Analysis

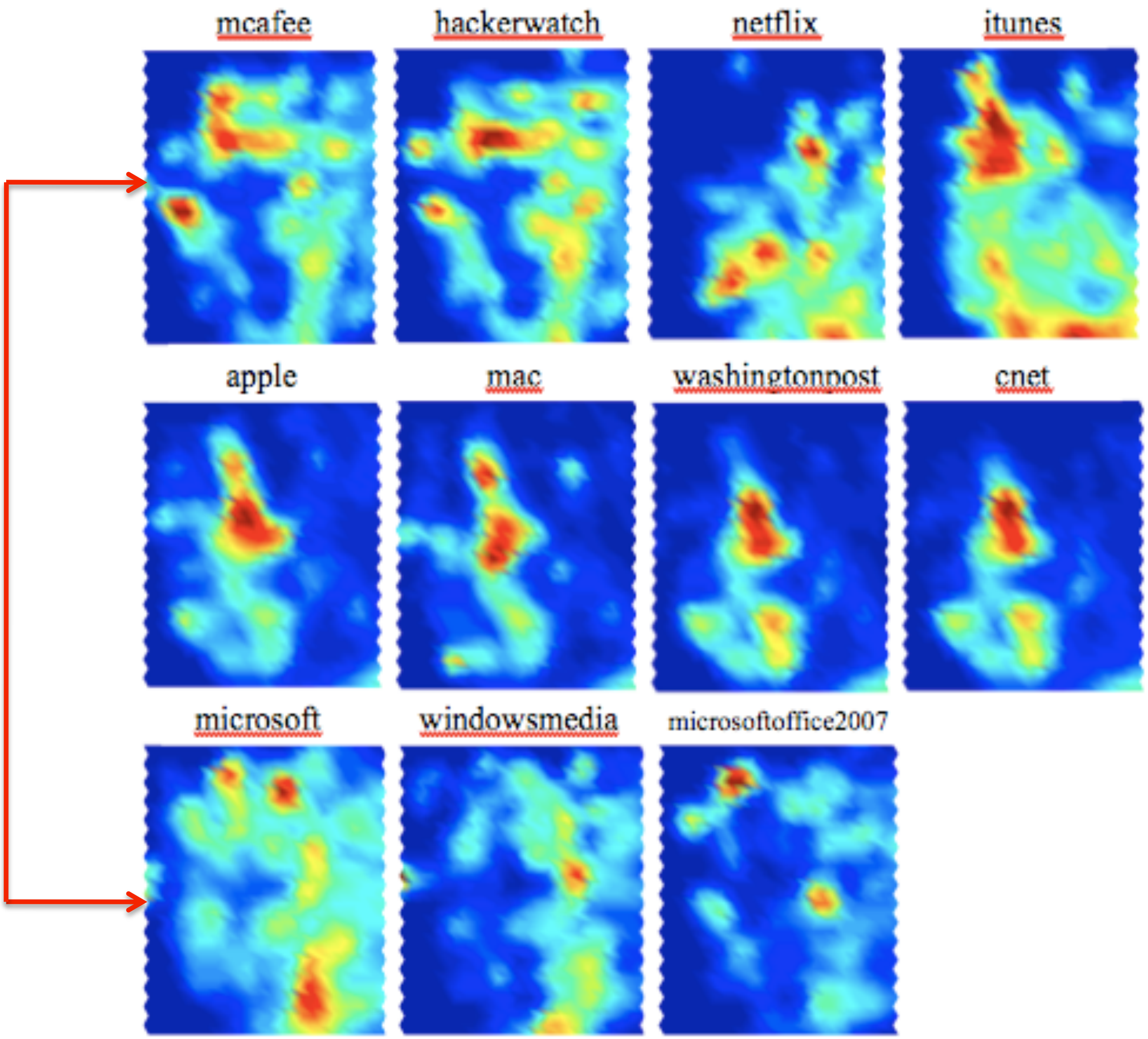
- Scale is still an issue
  - Investigate other, more efficient techniques
  - Add (more) multi-dimensional processing capability
  - Investigate scalable data mining systems [on-going]
- Meaningful visualization of results in multi-dimensions
- Behavior may not lend itself to one cluster
  - Need flexible, fuzzy clustering
- Centralized algorithms require global knowledge
  - Not fit for distributed implementation, protocols
  - Need more distributed, localized algorithms

Self-organizing Maps (SoMs)

- The topology-preserving mapping keeps the more similar data groups closer together in the final map



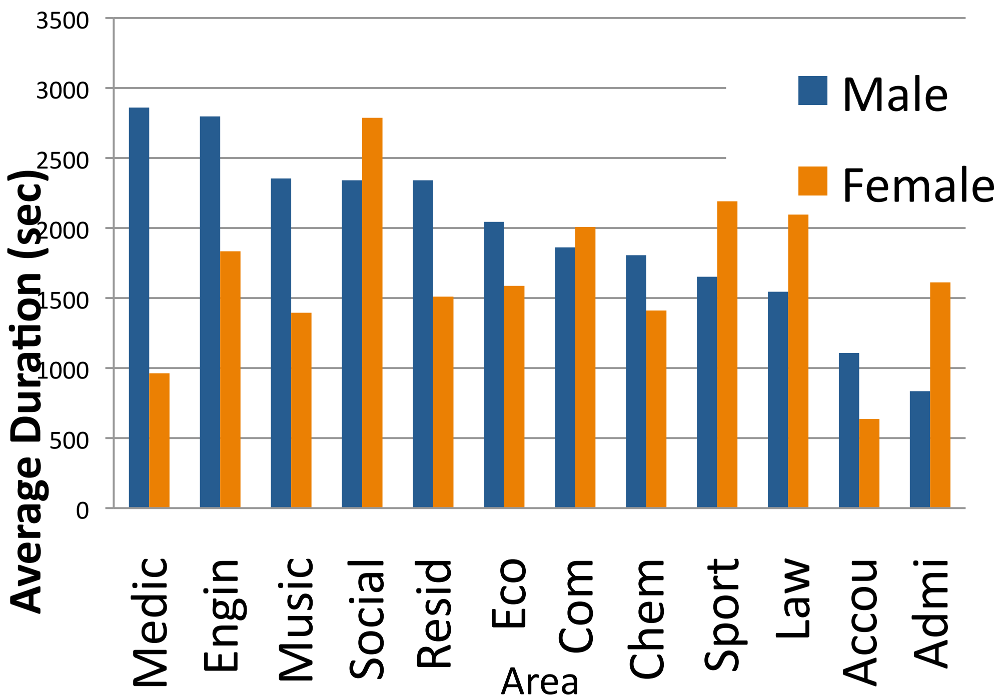
* S. Moghaddam, A. Helmy, "Internet Usage Modeling of Large Wireless Networks Using Self-Organizing Maps", *IEEE SCENES (MASS workshop)*, Nov 2010

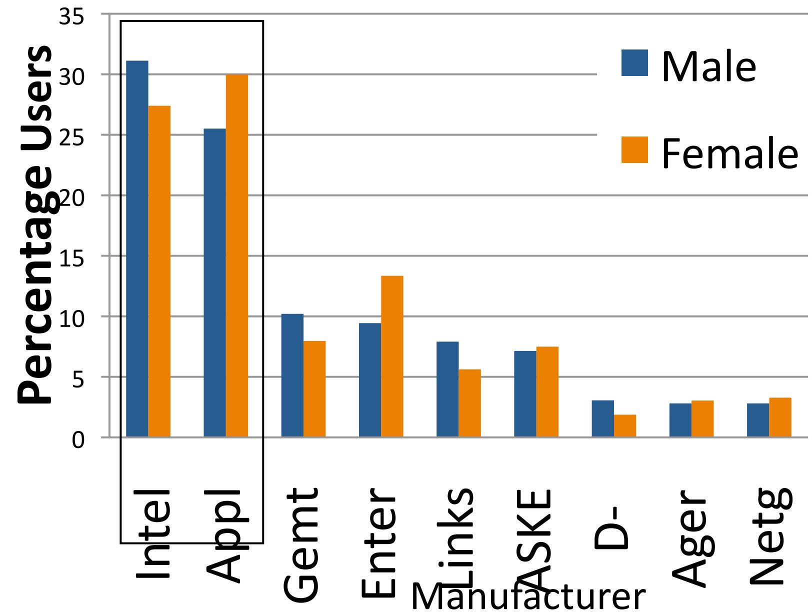


Feature maps for selected web domains

Gender-based feature analysis in Campus-wide WLANs

U. Kumar, A. Helmy, ACM MSWIM 2010 [SRC MobiCom 2007]





- Able to classify users by gender using knowledge of campus map

-Users exhibit distinct on-line behavior, preference of device and mobility based on gender

-On-going Work

  -How much more can we know?

  -What is the *"information-privacy trade-off"*?

Future Directions (Applications)

- Behavior aware push/caching services (targeted ads, events of interest, announcements)
- Caching based on behavioral prediction
- Detecting abnormal user behavior & access patterns based on previous profiles
- Can we extend this paradigm to include social aspects (trust, friendship, cooperation)?
- Privacy issues and mobile *k-anonymity*
- Participatory sensing, deputizing the community

The Need for *Trust* in Mobile Networks

- There is a need for cooperation and trust in mobile networks
- Peer-to-peer networks get formed using cooperation
- Without trust, there will be no network over which to run credit or reputation based systems
- Need to bootstrap trust in mobile networks. Trust can also be used in continued operation

U. Kumar, G. Thakur, A. Helmy, "*PROTECT: Proximity-based Trust-advisor using Encounters for Mobile Societies*", ACM IWCMC 2010, Journal under submission

Challenges and Promise

- Perception of lack of security (hence low trust)
  - Tetherless/wireless operation, - Mobility, dynamic topology
  - Lack of boundaries (firewalls/gateways) - Infrastructureless-ness
- Opportunities with encounters
  - Proximity, locality (spatial radio connectivity)
  - Encounter-based key establishment using out-of-band info or challenges
- Opportunities with behavioral modeling
  - Tight coupling between devices and users facilitates behavioral modeling
  - Can behavioral similarity be related to trust?

Establishment of Trust Advisors

- We propose the use of encounters and behavioral metrics to design ‘Trust filters’
- We investigate the characteristics of such metrics based on mobile network traces
- Are we likely to trust people similar to us? Social sciences [Homophily] suggest so
- Can we capitalize on behavioral similarity!

Trust Adviser Filters

- Frequency (or *Duration*) of *Encounter*: *FE* (or *DE*)
  - Order nodes based on freq. (or duration) of encounters. Pick top *T*% users.
- Behavior Vector (*BV*)
  - Based on duration or count of sessions
  - Inner product provides similarity score
- Behavior Matrix (*BM*)
  - Get the Eigen-Behaviors using (SVD)
  - Calculate users behavioral distance
  - Order nodes based on distance

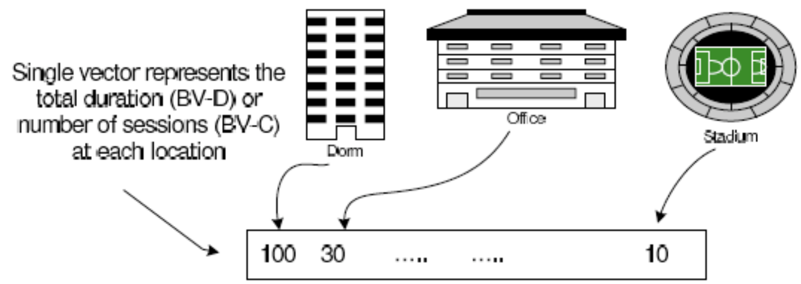


Figure 1: Behavior Vector for a user

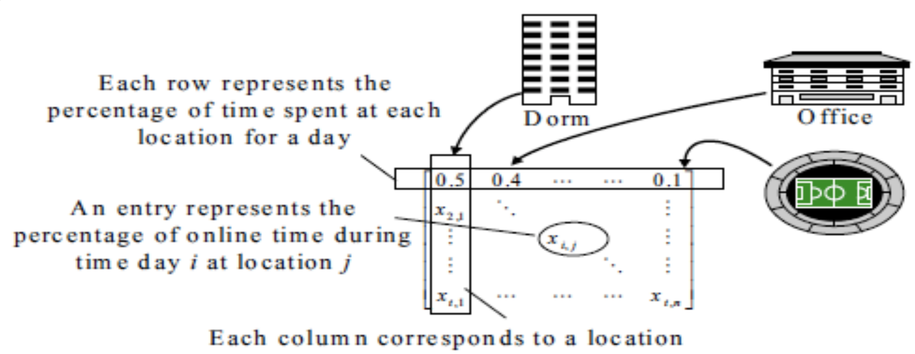
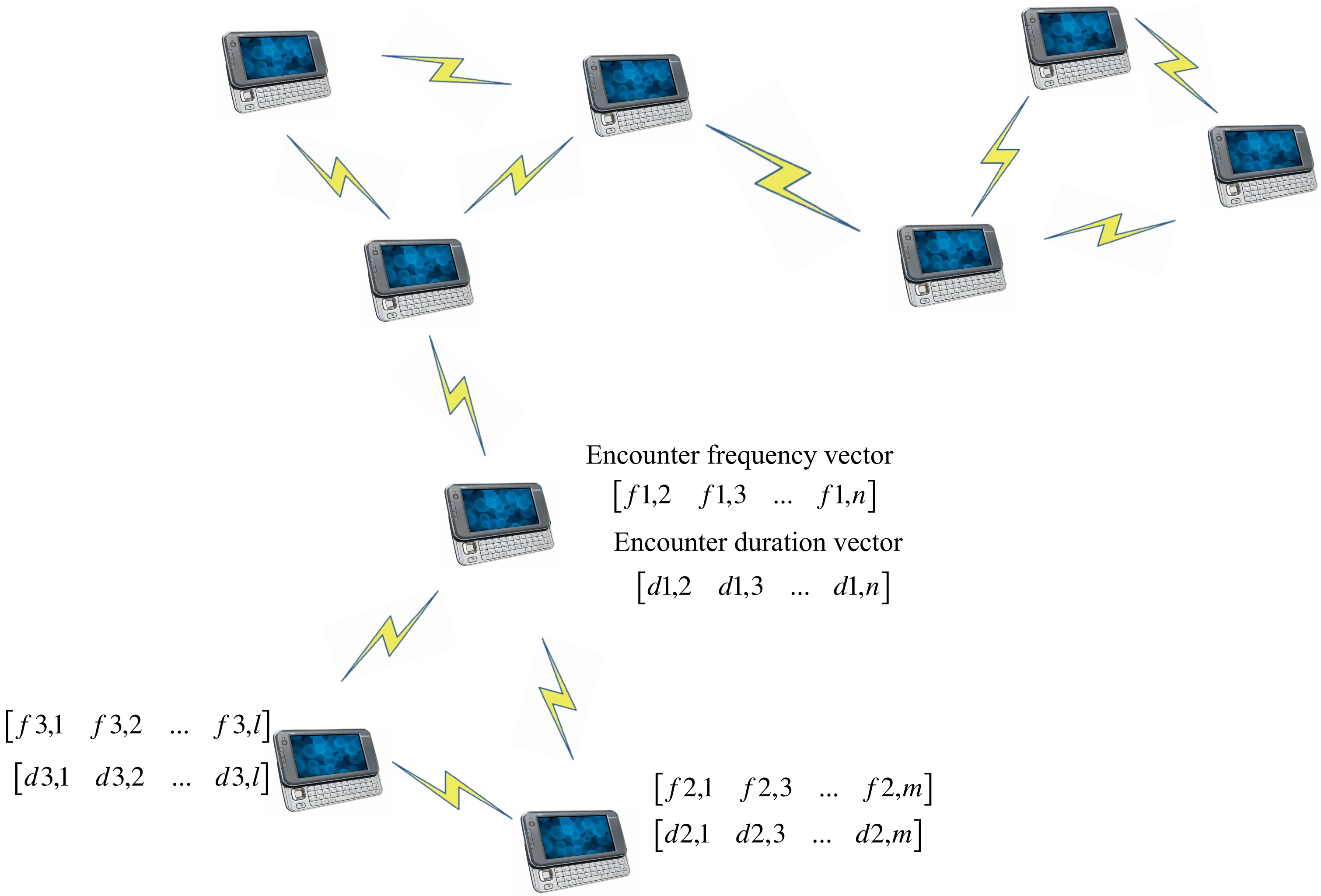
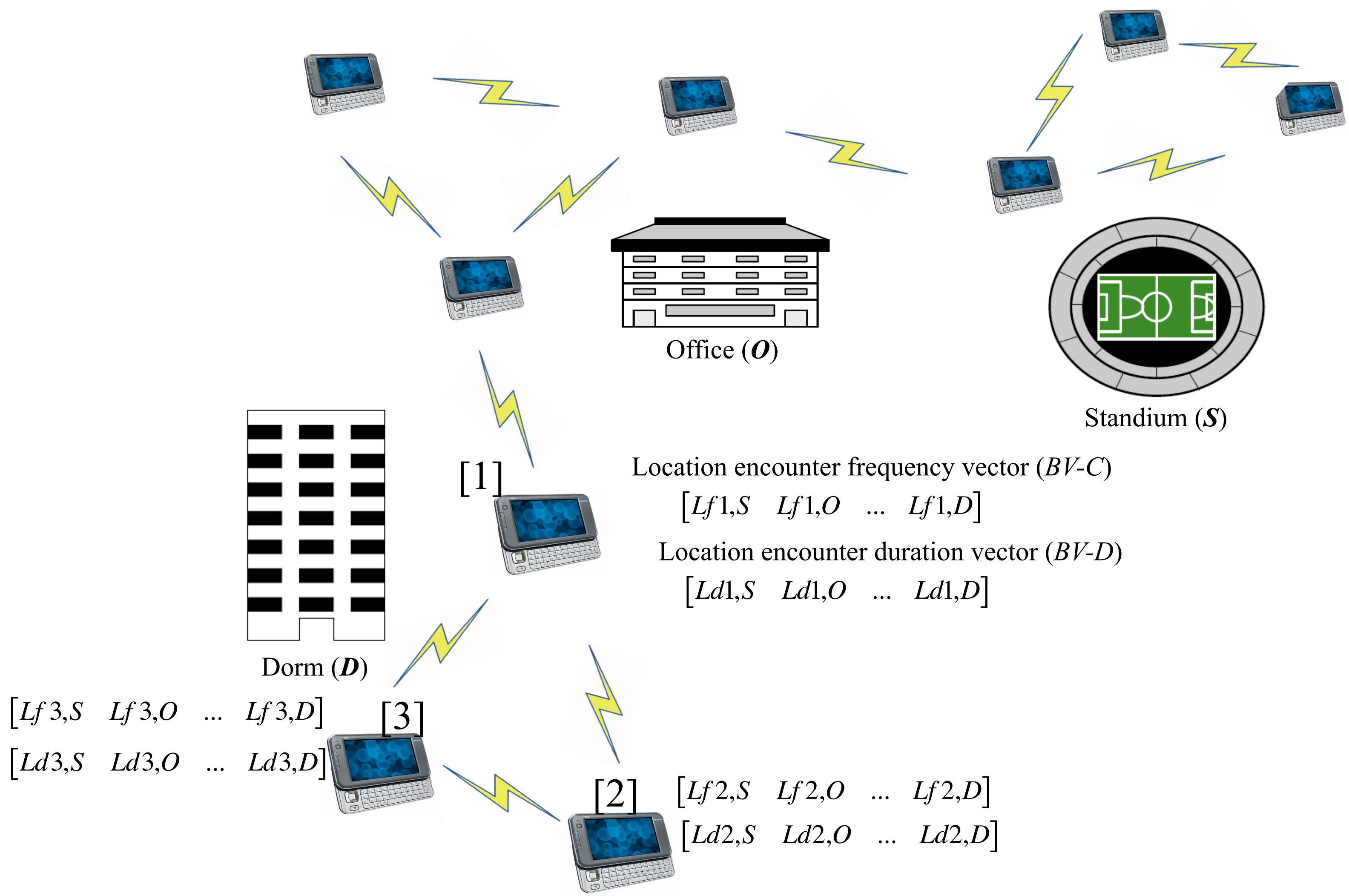


Figure 2: Behavior Matrix for a user



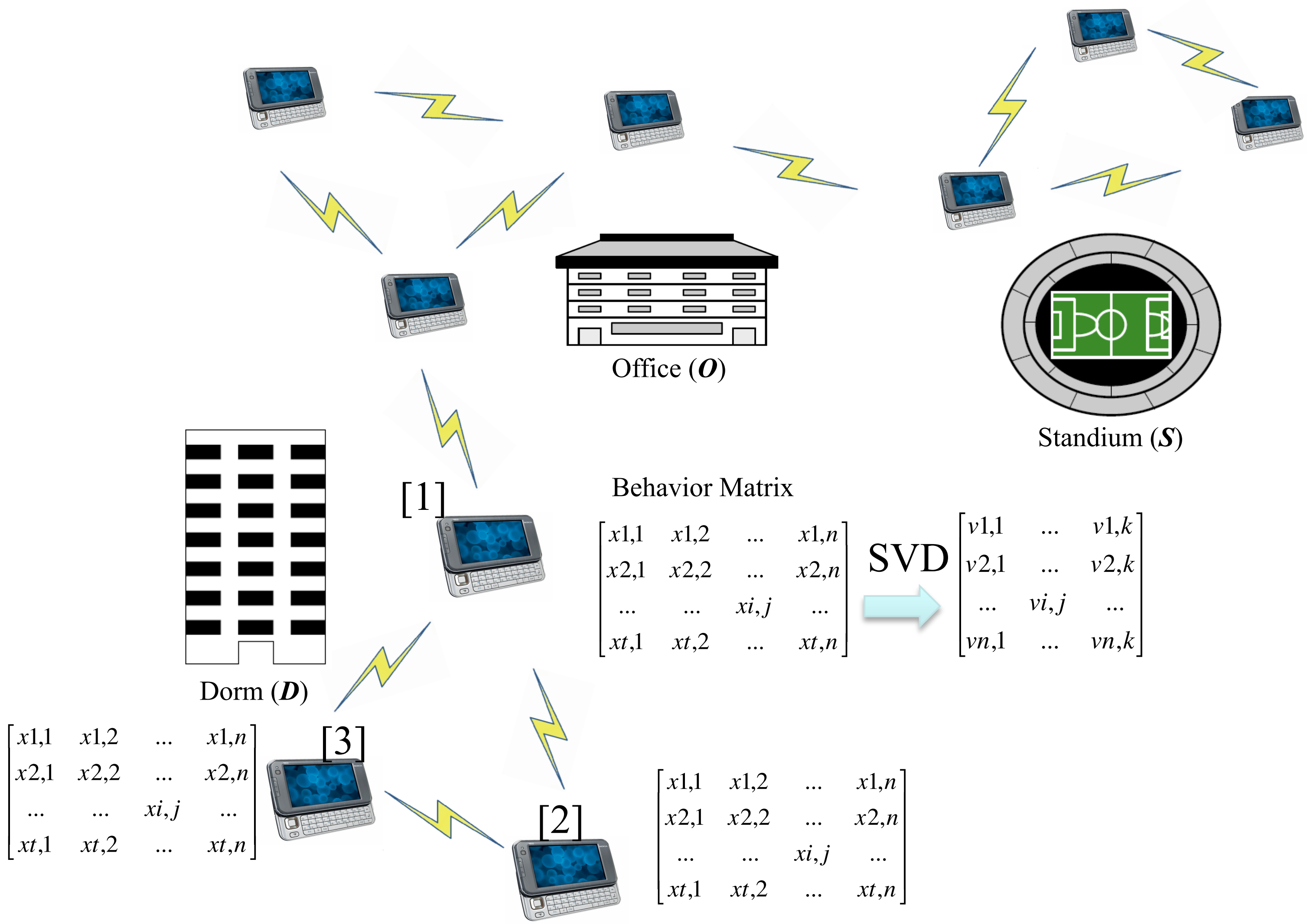
Devices construct encounter vectors based on frequency or duration of encounters with other nodes. In this example, node 1 encounters n other nodes, node 2 encounters m other nodes, and node 3 encounters l other nodes over the time frame of interest.

Frequency of Encounter and Duration of Encounter Trust Adviser Filters



Devices construct behavioral vectors based on location visitation patterns, including session frequency/count (*BV-C*) and duration of on-line activity (*BV-D*) at each location. In this example, devices visit three locations: the standium (***S***), the office (***O***), and the dorm (***D***).

Behavior Vector Trust Adviser Filters



Devices construct behavioral matrices based on location visitation patterns. SVD is used to obtain a summary of the major visitation trends and used for similarity comparisons.

Behavior Matrix Trust Adviser Filters

Evaluation and Analysis

- 1- Can the filters distinguish between different users? Statistical characterization of the encounter trends in the traces for the filters
- 2- Are the filters stable? how do trust lists change over time for each filter?
- 3- Can we achieve meaningful, stable trust (in Adhoc Nets) without sacrificing performance?

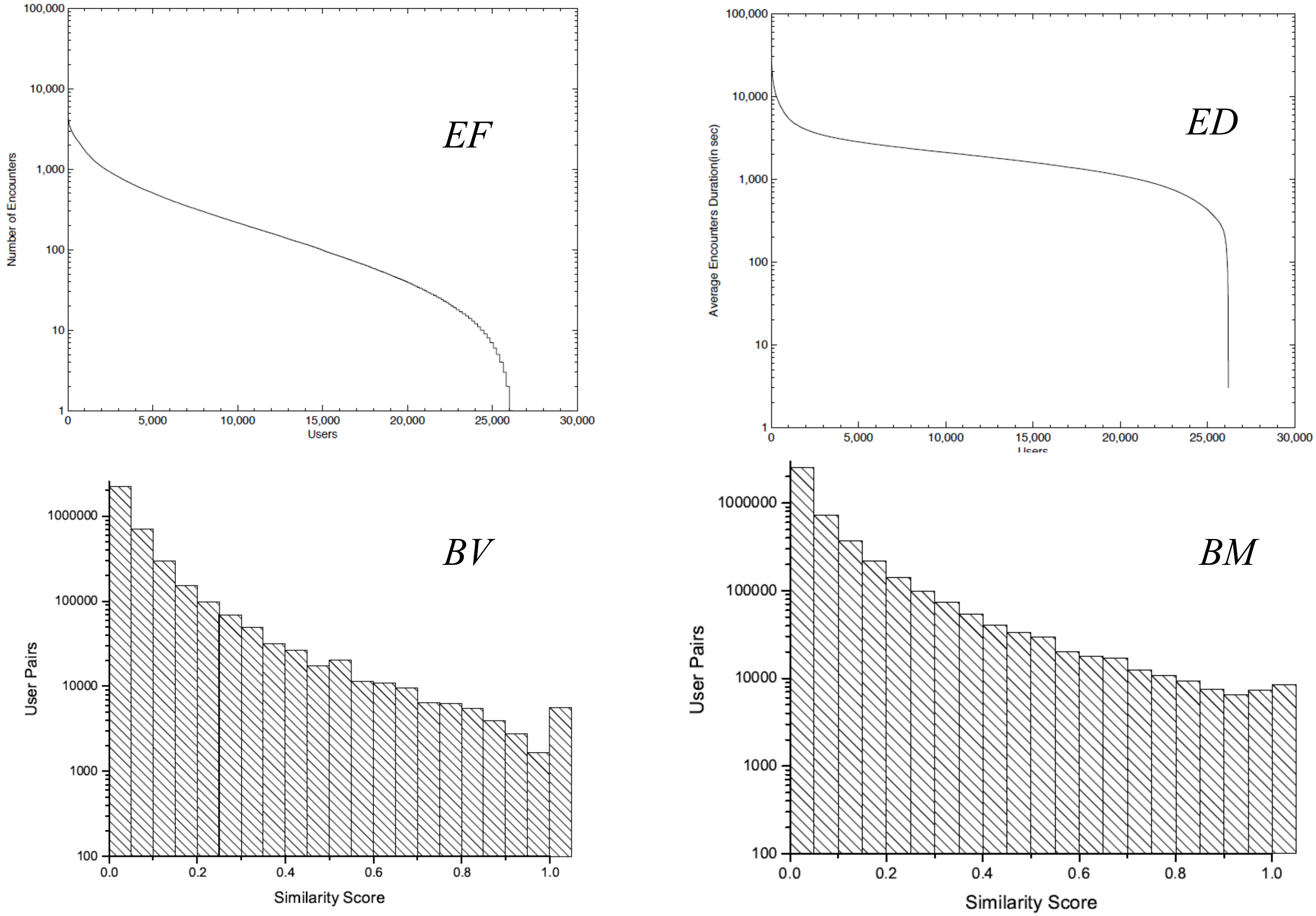
Traces Used

- 3 month long (Sep to Nov 2007) Wireless LAN (WLAN) traces from University of Florida, Gainesville.
- More than 35,000 users
- Total number of Access Points is over 730

MAC	Start Time	Duration(s)	AP/Location
00:11:22:33:44:55	01 Nov 2007 21:00:51 GMT	3000	CS_AP1
11:22:33:44:55:66	01 Nov 2007 21:01:30 GMT	10	ECE_AP2
01:02:03:04:05:06	01 Nov 2007 22:11:00 GMT	200	MSL_AP1
10:20:30:40:50:60	01 Nov 2007 22:15:30 GMT	600	MACA_AP1
11:22:33:44:55:66	01 Nov 2007 22:23:10 GMT	180	CS_AP3

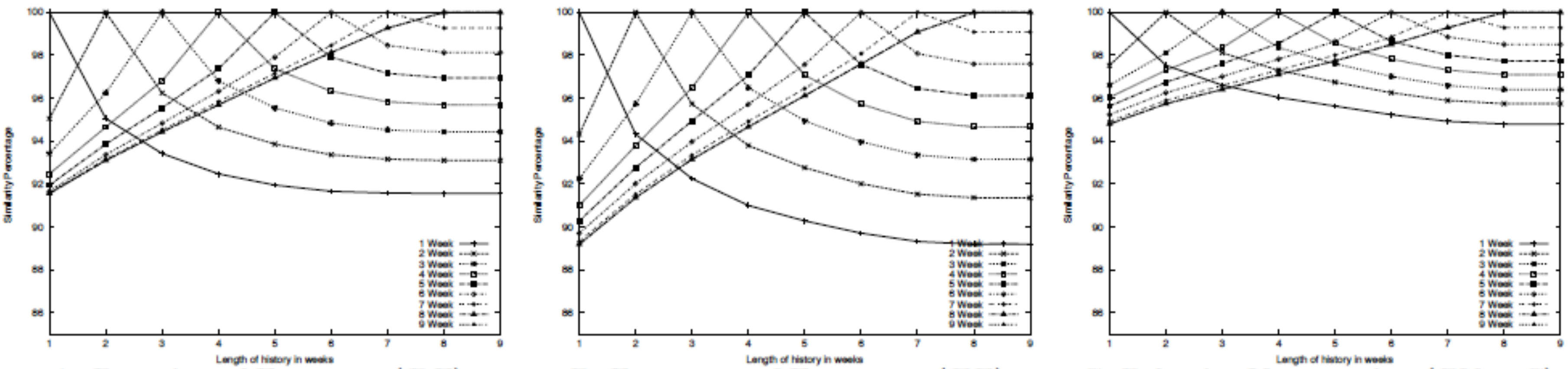
Table 1: Sample WLAN trace

Characterization of Encounter Stats with Trust Filters



- Richness of encounter distributions (and the *knee*) could differentiate users

Filter Stability Analysis

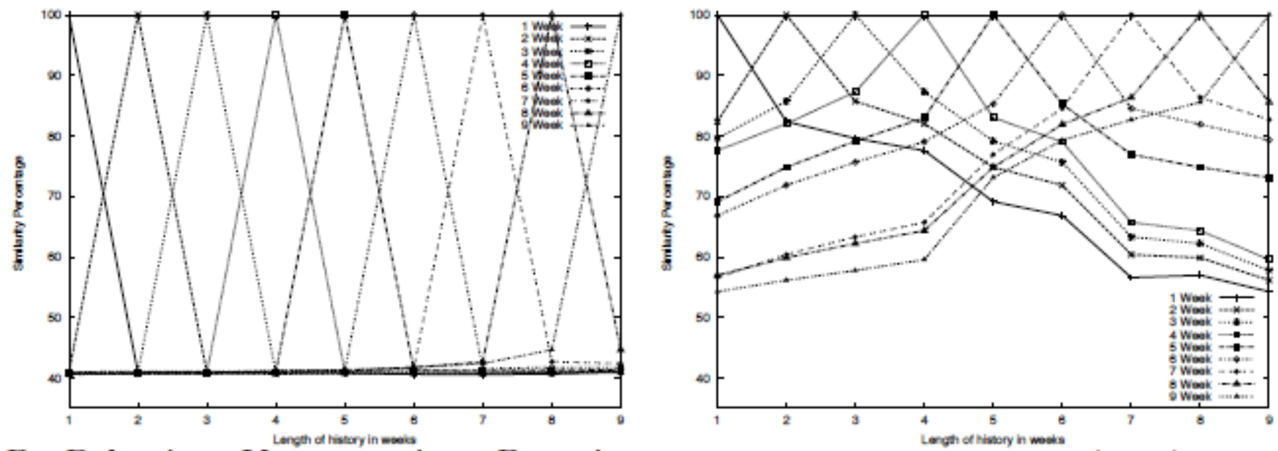


A. Duration of Encounter(DE) B. Frequency of Encounter (FE) C. Behavior Vector using ($BV - C$)

Figure 7: Comparison of trust list belonging to different history for various filters at T=40% (note that the y-axis scale for DE, FE, and $BV - C$ starts at 85% and for $BV - D$ and BM the scale starts at 35%)

- Desirable to possess stability in the advisory lists over time
- Behavior vector based on session count (*BV-C*) filter is the most stable with over 95% over 9 weeks
- Freq. (*FE*) and duration of encounter (*DE*) filters have good stability with over 89% common users over 9 weeks

Filter Stability Analysis (contd.)



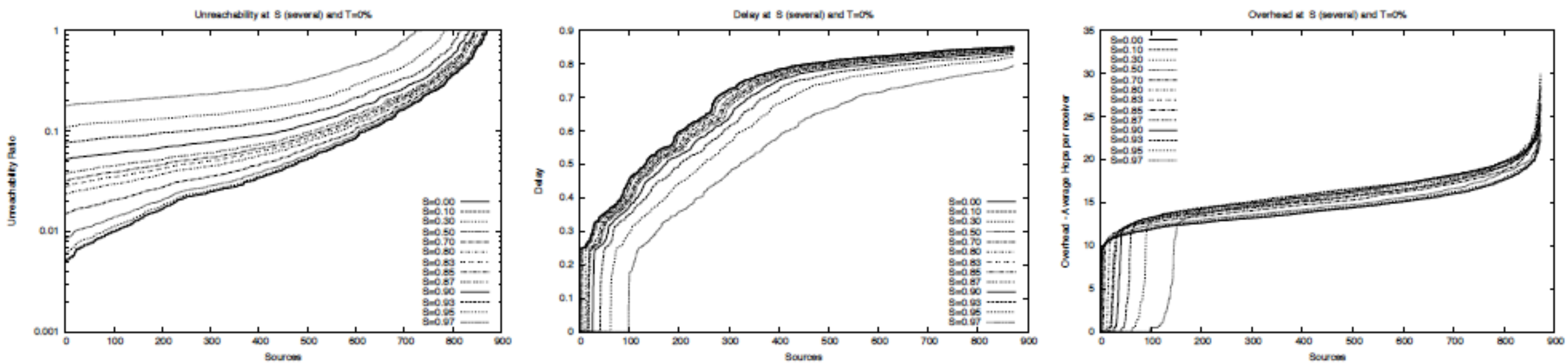
D. Behavior Vector using Duration ($BV - D$)

E. Behavior Matrix (BM)

Figure 7: Comparison of trust list belonging to different history for various filters at T=40% (note that the y-axis scale for DE, FE, and $BV - C$ starts at 85% and for $BV - D$ and BM the scale starts at 35%)

- Behavior vector based on duration (*BV-D*) is the least stable with ~40% stability over 1-9 weeks
- Behavior matrix is relatively stable (~80%) for 3 weeks. Stability degrades to ~55% for 9 wks

Epidemic Routing Analysis with Selfishness (no Trust)



A. Unreachability B. Delay C. Overhead

Figure 9: Unreachability, Delay and Overhead at several S and T=0%, during trace period of Nov 2007

- Reachability degrades noticeably with increased selfishness
- DTN routing suffers significantly with selfishness
- Can trust help?

Epidemic Routing with Selfishness and Trust

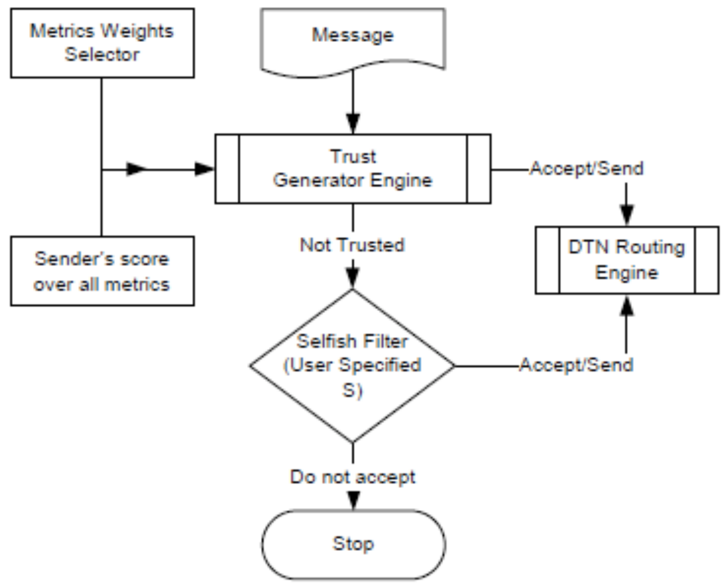


Figure 10: Flow chart for making routing decisions

- Trust-augmented DTN routing engine
  - If the sending node is trusted (according to a trust adviser filter) then accept and forward message
  - Otherwise, do not forward if selfish to sender

Epidemic Routing Analysis with Selfishness (with Trust)

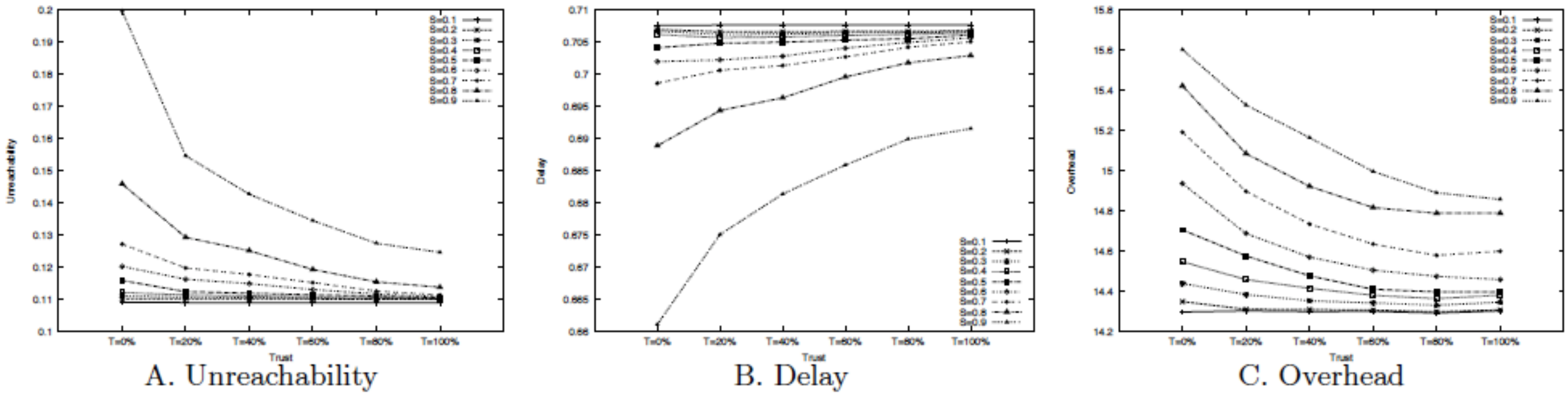


Figure 12: Average Unreachability, Delay and Overhead at various values of T and S for *DE* filter

- Q: Can we use trust without much sacrifice to performance?
- A: Trust can be used with selective choice of nodes without losing on performance. Enhancing performance over selfish cases dramatically

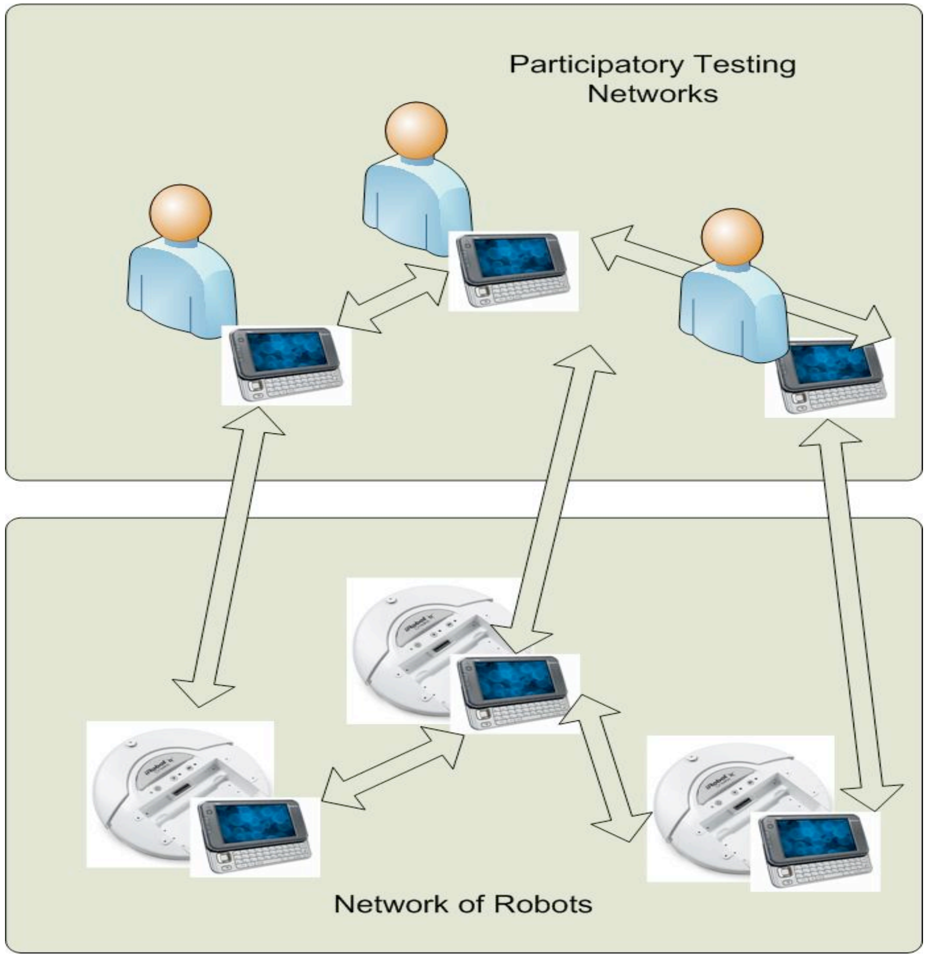
Conclusions

- Richness of encounter and behavioral patterns provide promise to differentiate between users
- Trust advisors can selectively choose nodes (in a stable, meaningful way) without sacrificing routing performance

Future Work

- Add location and context of encounters
- Evaluate with other traces and DTN routing
- Development of full trust-based protocol
- Validation: *surveys*, implementation & deployment

Participatory Sensing: Test beds with an Attitude



* S. Moon, A. Helmy, “Mobile Test beds with an Attitude!”, ACM MobiCom, ACM WinTECH [demo competitions], Sep 2010

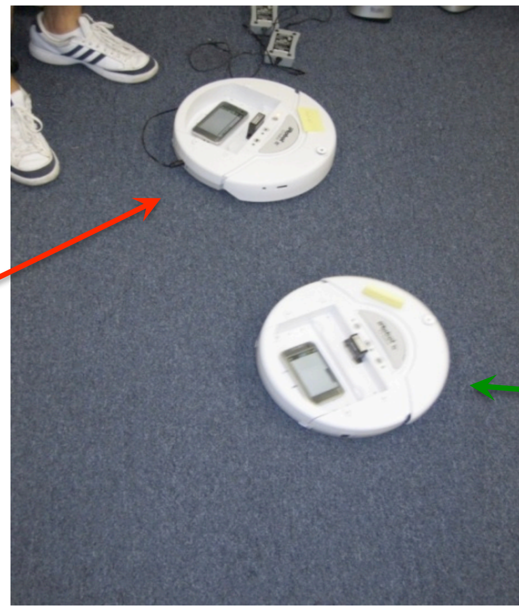
- Building personalities (i.e., attitudes) for mobile nodes

Behavior Matrix *M2*

$$\begin{bmatrix} x1,1 & x1,2 & \dots & x1,n \\ x2,1 & x2,2 & \dots & x2,n \\ \dots & \dots & xi,j & \dots \\ xt,1 & xt,2 & \dots & xt,n \end{bmatrix} \xrightarrow{\text{SVD}} \begin{bmatrix} v1,1 & \dots & v1,k \\ v2,1 & \dots & v2,k \\ \dots & vi,j & \dots \\ vn,1 & \dots & vn,k \end{bmatrix}$$

Behavior Matrix *M1*

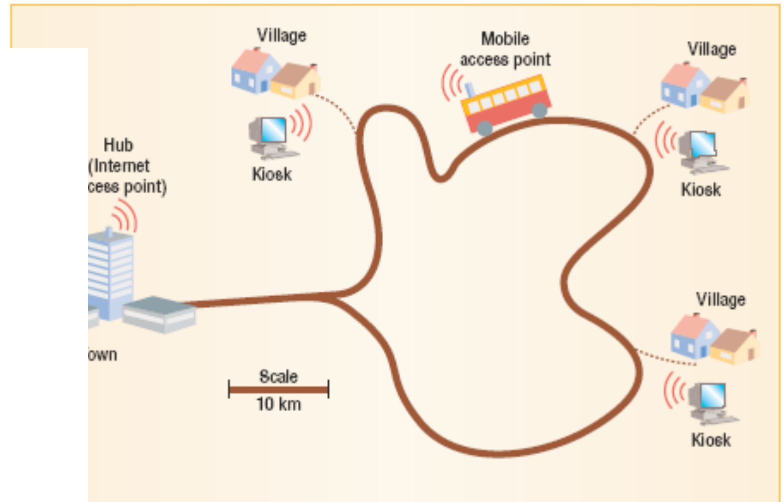
$$\begin{bmatrix} x1,1 & x1,2 & \dots & x1,n \\ x2,1 & x2,2 & \dots & x2,n \\ \dots & \dots & xi,j & \dots \\ xt,1 & xt,2 & \dots & xt,n \end{bmatrix} \xrightarrow{\text{SVD}} \begin{bmatrix} v1,1 & \dots & v1,k \\ v2,1 & \dots & v2,k \\ \dots & vi,j & \dots \\ vn,1 & \dots & vn,k \end{bmatrix}$$



Building communities of autonomous mobile nodes (e.g., robots), that can interact with each other and with communities of participatory testers to form friendship, trust, interest …

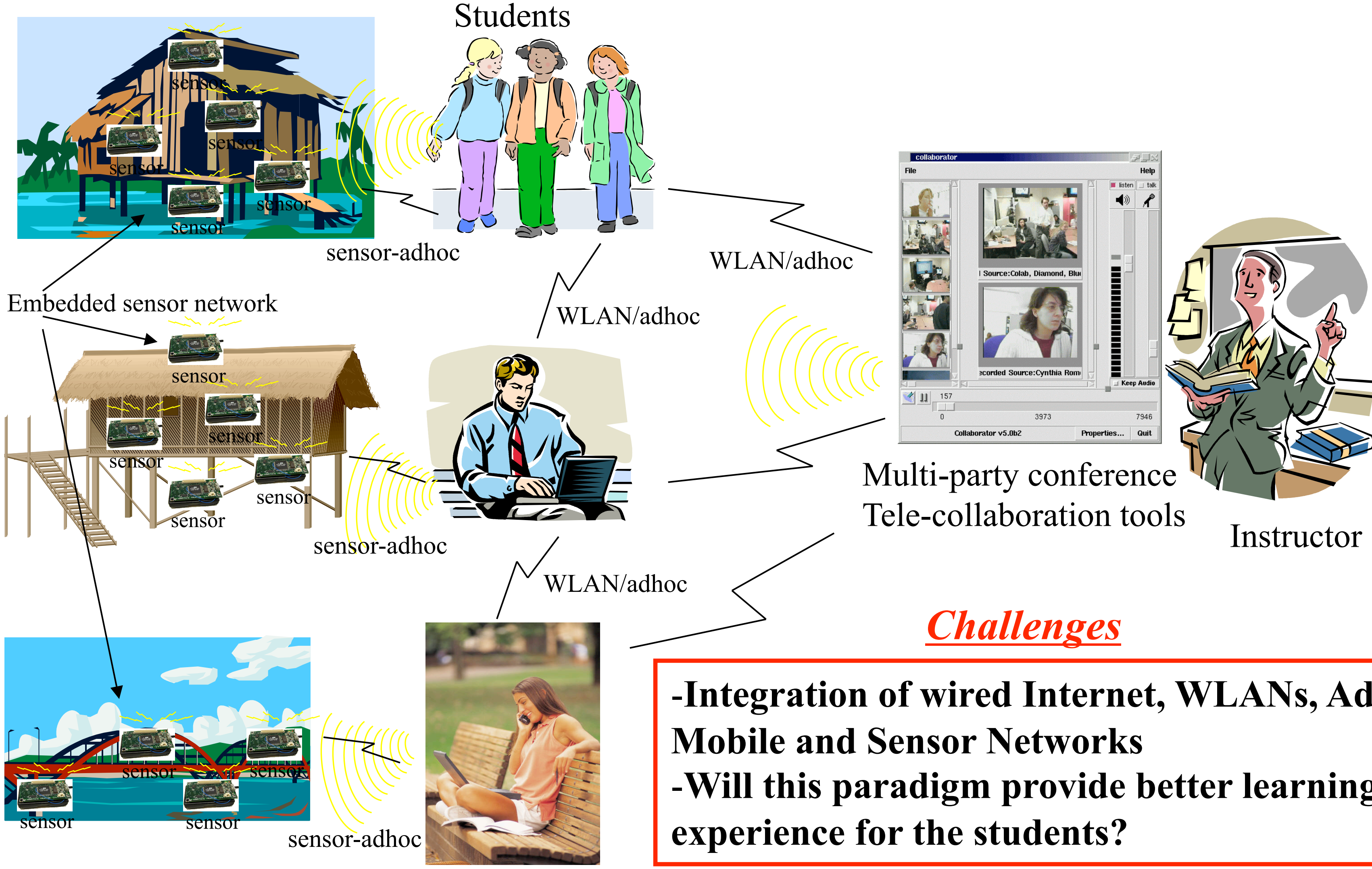
Disaster Relief (Self-Configuring) Networks

On-going and Future Directions Utilizing mobility



- Controlled mobility scenarios
  - DakNet, Message Ferries, Info Station
- Mobility-Assisted protocols
  - Mobility-assisted information diffusion: EASE, FRESH, DTN, $100 laptop
- Context-aware Networking
  - Mobility-aware protocols: self-configuring, mobility-adaptive protocols
  - Socially-aware protocols: security, trust, friendship, associations, small worlds
- On-going Projects
  - Next Generation (Boundless) Classroom
  - Disaster Relief Self-configuring Survivable Networks

The Next Generation (Boundless) Classroom



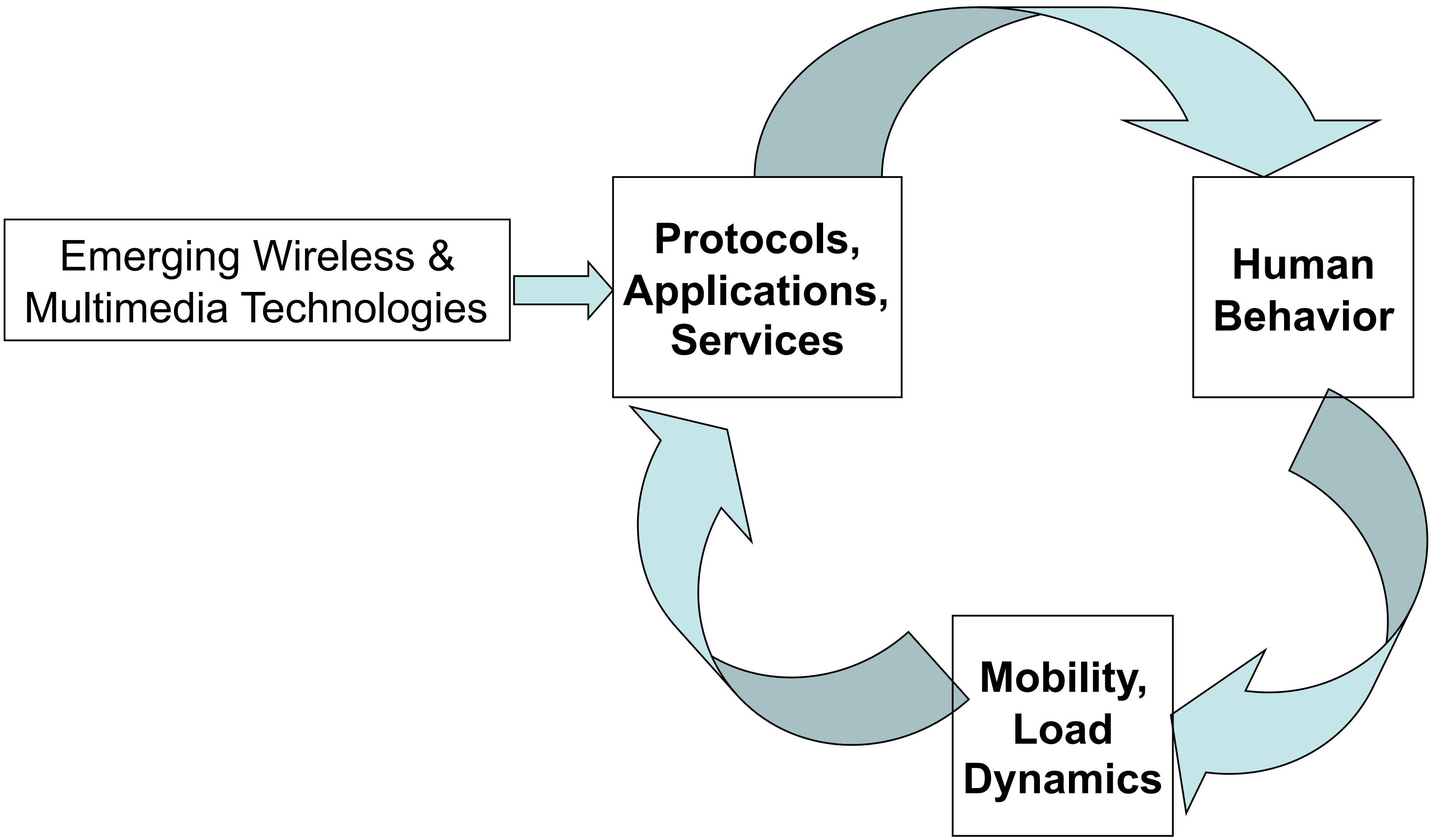
Challenges

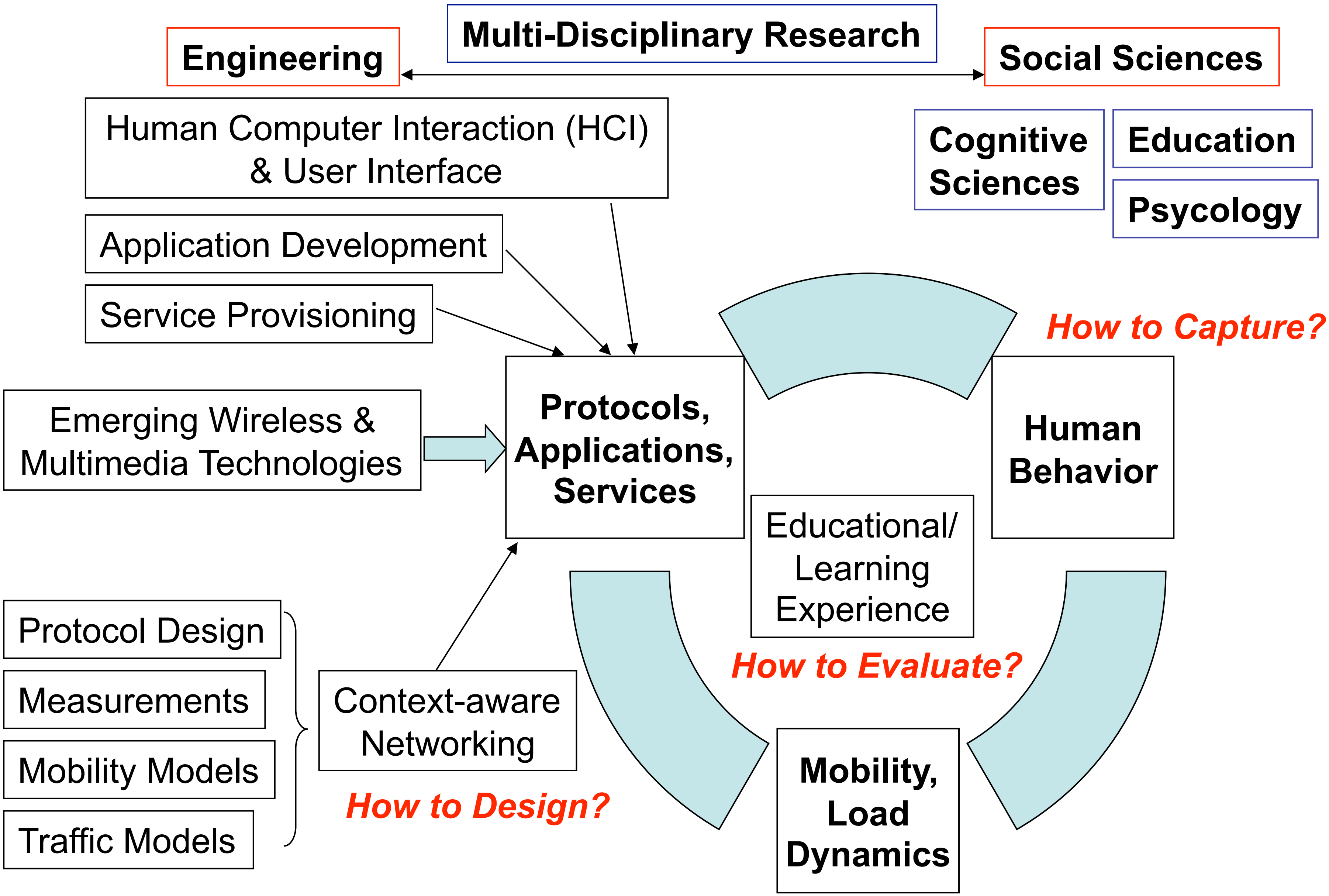
-Integration of wired Internet, WLANs, Adhoc Mobile and Sensor Networks
-Will this paradigm provide better learning experience for the students?

Real world group experiments (structural health monitoring)

Future Directions: Technology-Human Interaction

The Next Generation Classroom




Multi-Disciplinary Research
Engineering
Social Sciences
Human Computer Interaction (HCI) & User Interface
Application Development
Service Provisioning
Cognitive Sciences
Education
Psycology
How to Capture?
Emerging Wireless & Multimedia Technologies
Protocols, Applications, Services
Human Behavior
Educational/ Learning Experience
How to Evaluate?
Protocol Design
Measurements
Mobility Models
Traffic Models
Context-aware Networking
How to Design?
Mobility, Load Dynamics

Thank you!

Ahmed Helmy helmy@ufl.edu
URL: www.cise.ufl.edu/~helmy
MobiLib: nile.cise.ufl.edu/MobiLib