

Leveraging mobile network big data for development

Sriganesh Lokanathan, LIRNEasia

IIT-Delhi, New Delhi

22 August 2014



This work was carried out with the aid of a grant from the International Development Research Centre, Ottawa, Canada.

About LIRNEasia

(www.lirneasia.net)

- We are a regional think tank based in Colombo, working across South Asia, South East Asia and the Pacific Small Island States
- Our mission
 - Catalyzing policy change through research to improve people's lives in the emerging Asia Pacific by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology

What is Big Data?

Big Data

The Vs

VELOCITY

Rate of change
of data

VOLUME

Size of data

VARIETY

Different forms
of data

VERACITY

Uncertainty of
data

VALUE

Potential value
of data

What has facilitated the rise of big data?

- Vast drops in the cost of storing and retrieving information
- Exponential growth in computer power and memory
 - data can reside in persistent memory instead of disk and tape
- Major improvements in techniques for performing machine learning and reasoning

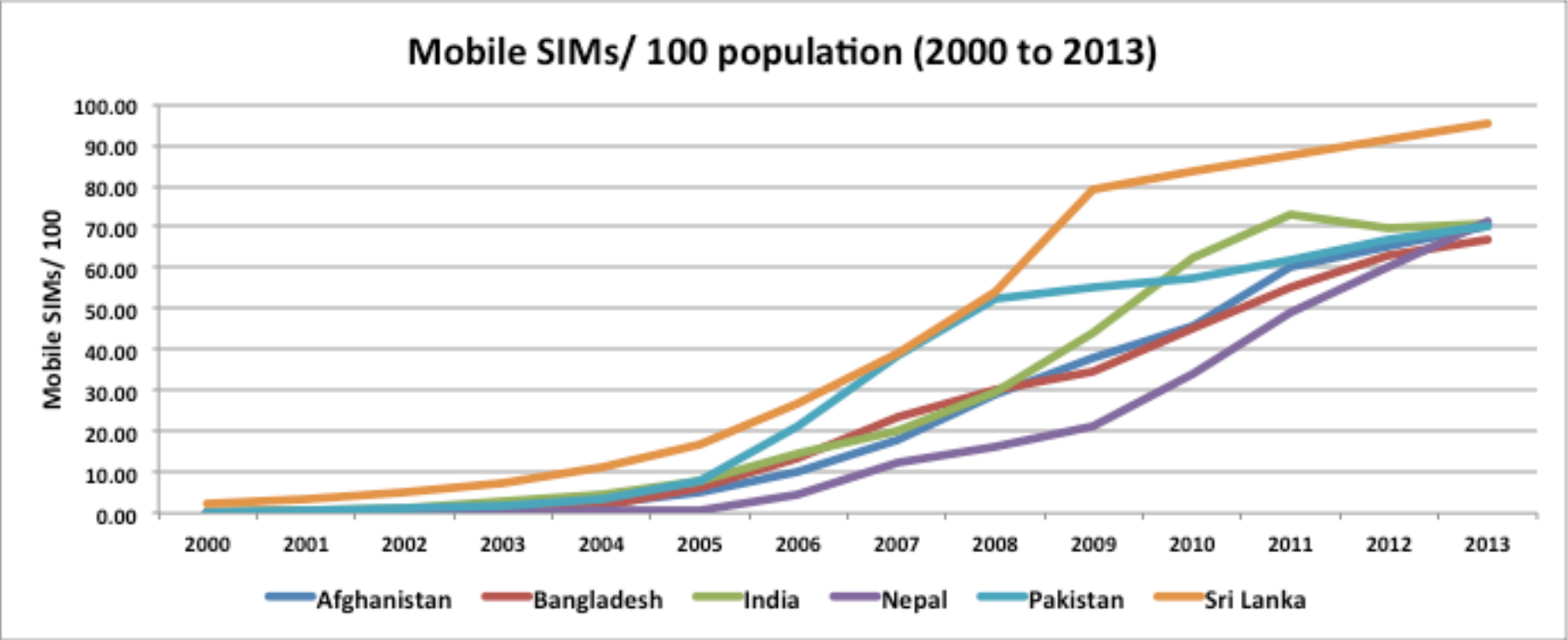
Sources of big data

- Administrative data
 - E.g. digitized medical records, insurance records, tax records, etc.
- Commercial transactions
 - E.g. Bank transactions, credit card purchases, supermarket purchases, online purchases, etc.
- Sensors and tracking devices
 - E.g. road and traffic sensors, climate sensors, equipment & infrastructure sensors, mobile phones, satellite/ GPS devices, etc.
- Online activities/ social media
 - E.g. online search activity, online page views, blogs/ FB/ twitter posts, online audio/ video/ images, etc.

If we want comprehensive coverage of the population, what are the sources of big data in developing economies?

- Administrative data?
 - E.g. digitized medical records, insurance records, tax records, etc.
- Commercial transactions?
 - E.g. Bank transactions, credit card purchases, supermarket purchases, online purchases, etc.
- Sensors and tracking devices?
 - E.g. road and traffic sensors, climate sensors, equipment & infrastructure sensors, mobile phones, satellite/ GPS devices, etc.
- Online activities/ social media?
 - E.g. online search activity, online page views, blogs/ FB/ twitter posts, online audio/ video/ images, etc.

Currently only mobile network big data has the widest possible population coverage



LIRNEasia's Big Data for Development (BD4D) Research

- LIRNEasia has negotiated access to **historical and anonymized** telecom network meta-data from multiple operators in Sri Lanka
- In the current research cycle we are
 - conducting exploratory research on answering a few social science questions related to mobility and connectedness
 - developing a framework with privacy and self-regulatory guidelines for the collection, use and sharing of mobile phone data.
- <http://lirneasia.net/projects/bd4d/>

The data sets

- Multiple mobile operators in Sri Lanka have provided LIRNEasia access to 4 different types of meta-data:
 - Call Detail Records (CDRs)
 - Records of calls, SMS-es, Internet access
 - Airtime top-up records
- Data sets do not include any Personally Identifiable Information (PII).
 - All phone numbers are anonymized and
 - LIRNEasia does not maintain any mappings of identifiers to original phone numbers

What are some types of big data captured by mobile network operators?

- Call Detail Record (CDR)
 - Records of all calls made and received by a person created mainly for the purposes of billing
 - Similar records exist for all SMS-es sent and received as well as for all Internet sessions

Calling Party Number	Called Party Number	Caller Cell ID	Call Time	Call Duration
A24BC1571X	B321SG141X	3134	13-04-2013 17:42:14	00:03:35

- The Cell ID in turn has a lat-long position associated with it.

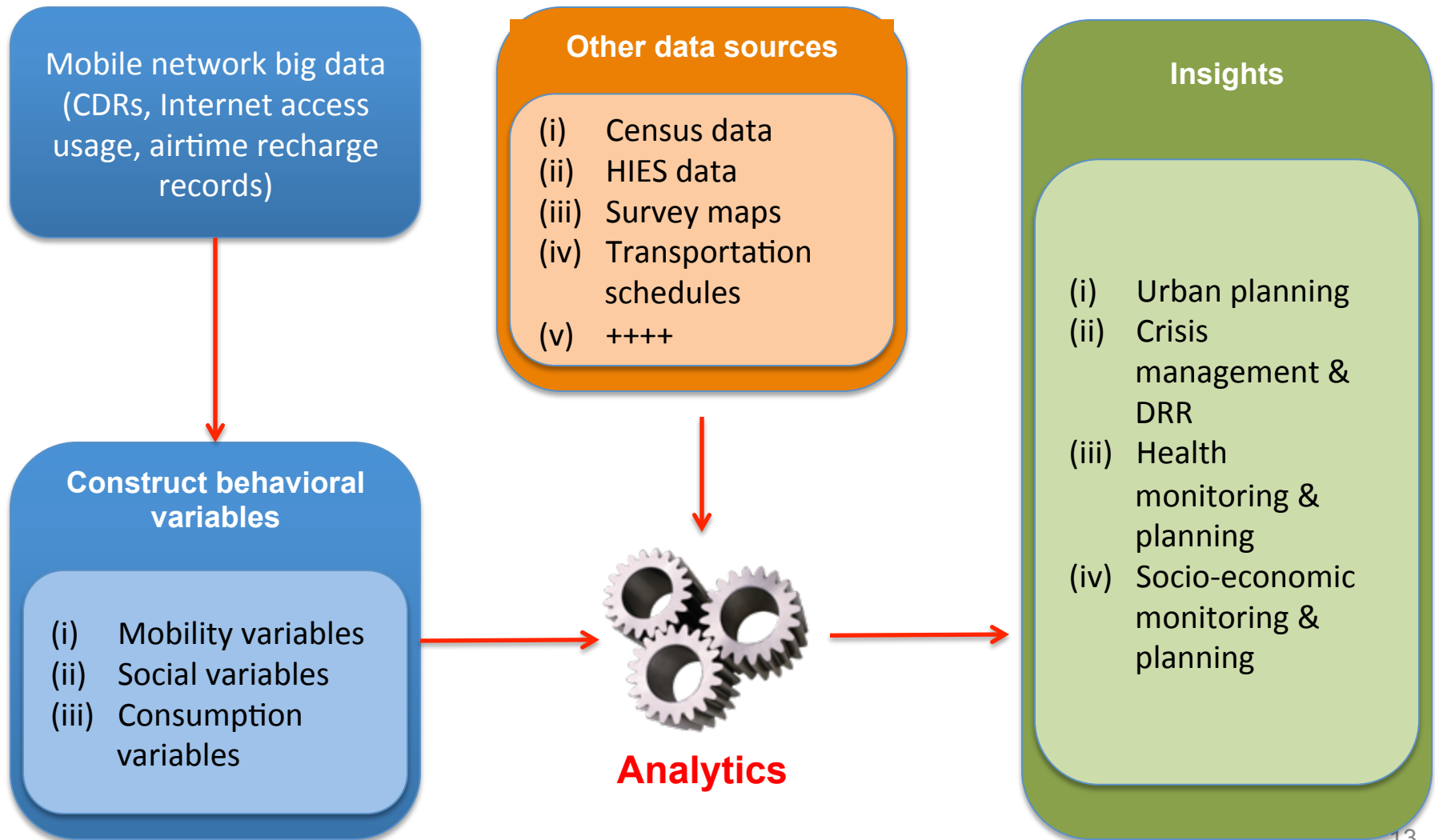
What are some types of big data captured by mobile network operators (contd.)?

- **Airtime reload records**

- Records of all airtime reloads performed by prepaid SIMs
- Each row corresponds to a record of one person's activity:

Number	Type of recharge	Starting balance	Amount	Time
A24BC1571X	CARD	0.41	50	13-04-2013 17:42:14

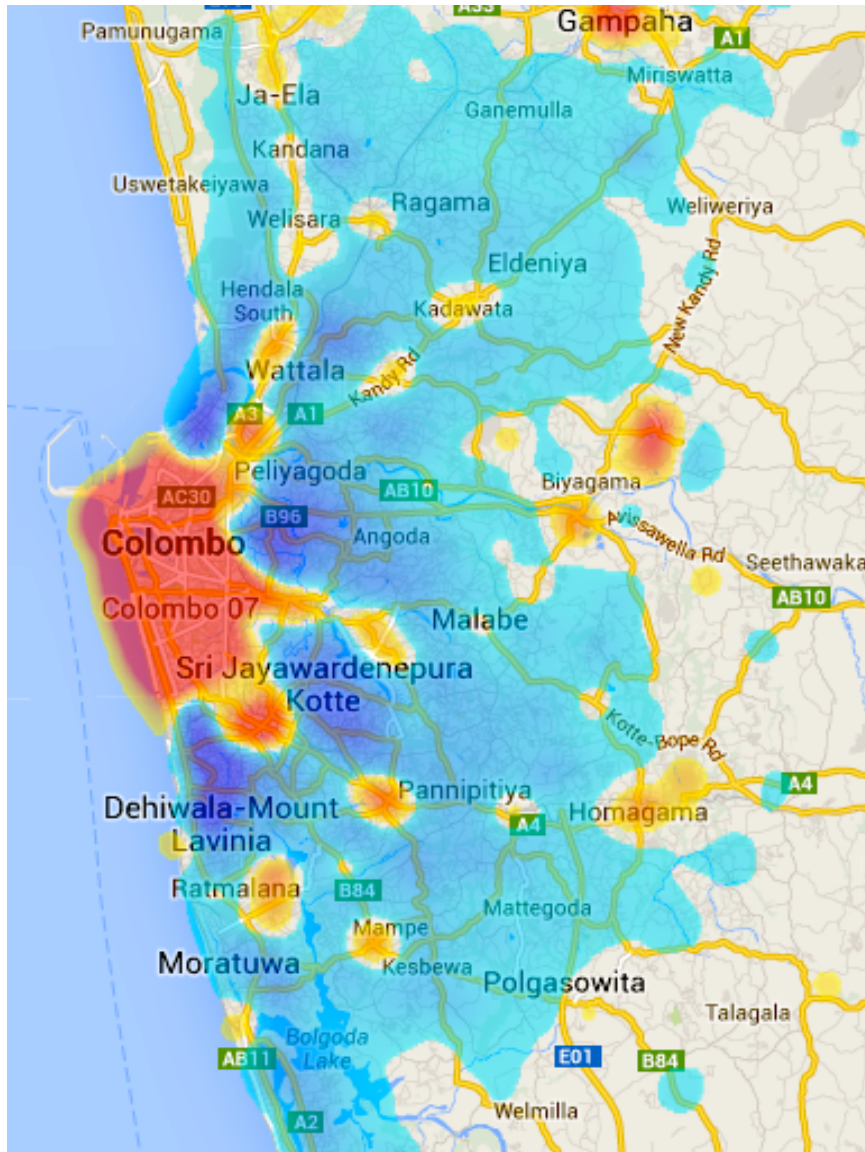
The overall process of leveraging mobile network big data for development



Why are we doing this work?

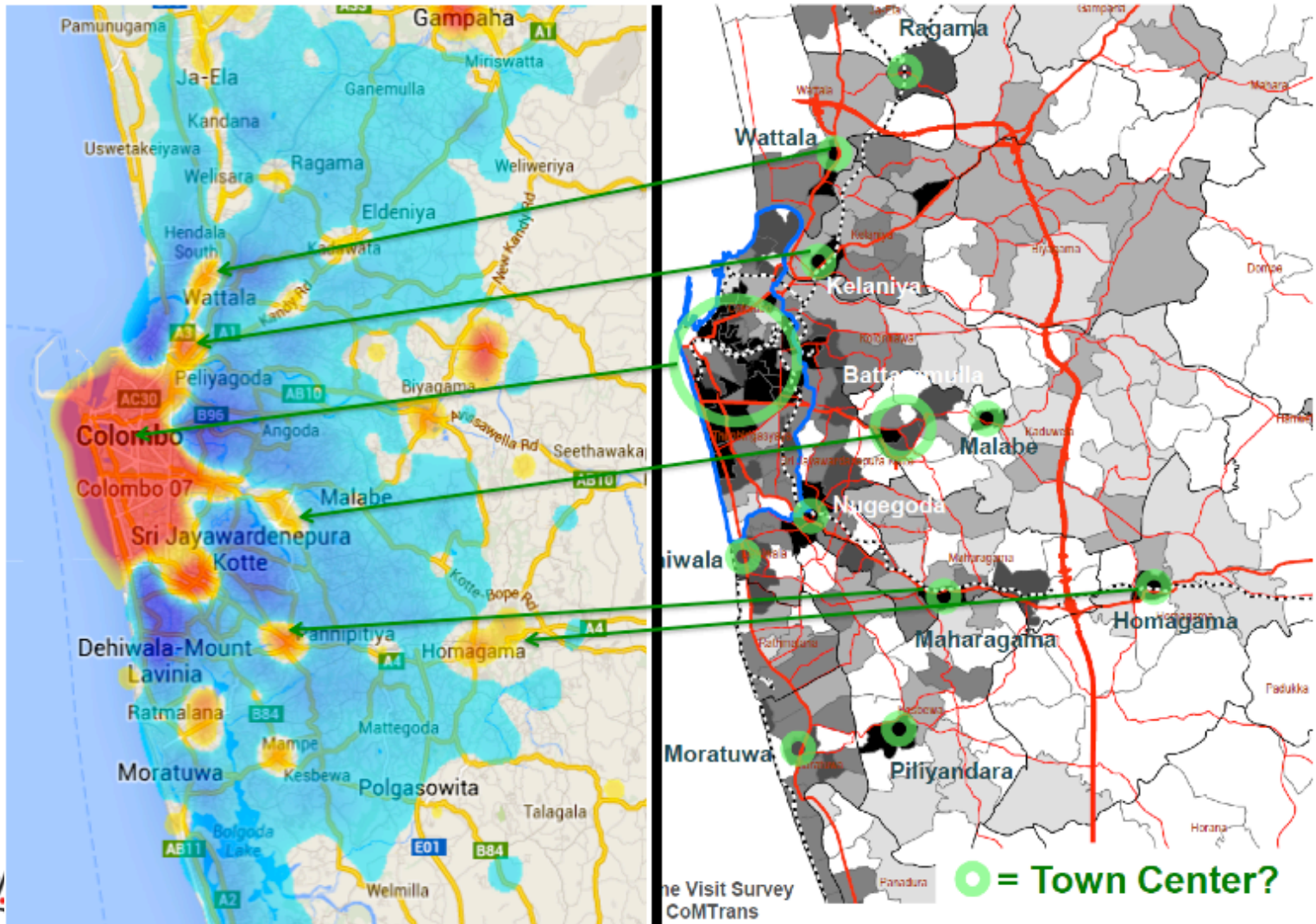
- To bring timely evidence into the policy making process in developing economies

An example of timely policy-relevant evidence



- The image on the left depicts relative density of people in Colombo city and the surrounding regions at 1300 compared to 0000 (midnight the previous day) on a normal weekday.
- The yellow to red colors depict areas whose density has increased relative to midnight. The blue color depicts areas whose density has decreased relative to midnight (the darker the blue, the greater the loss in density). The clear areas are those where the overall density has not changed.

Our findings closely match results from expensive & infrequent transportation surveys



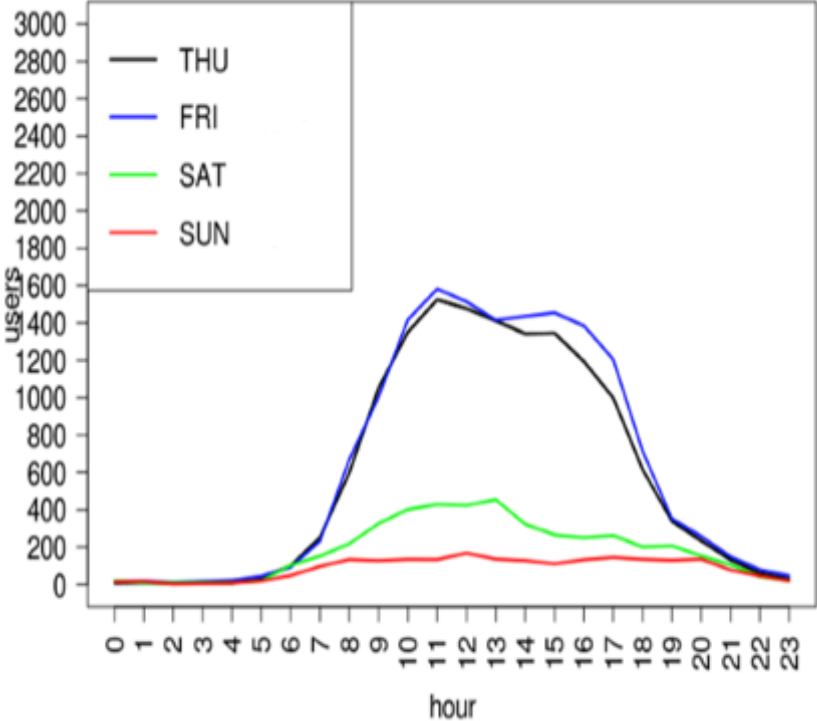
Using mobile network big data we can understand the mobility patterns of the population



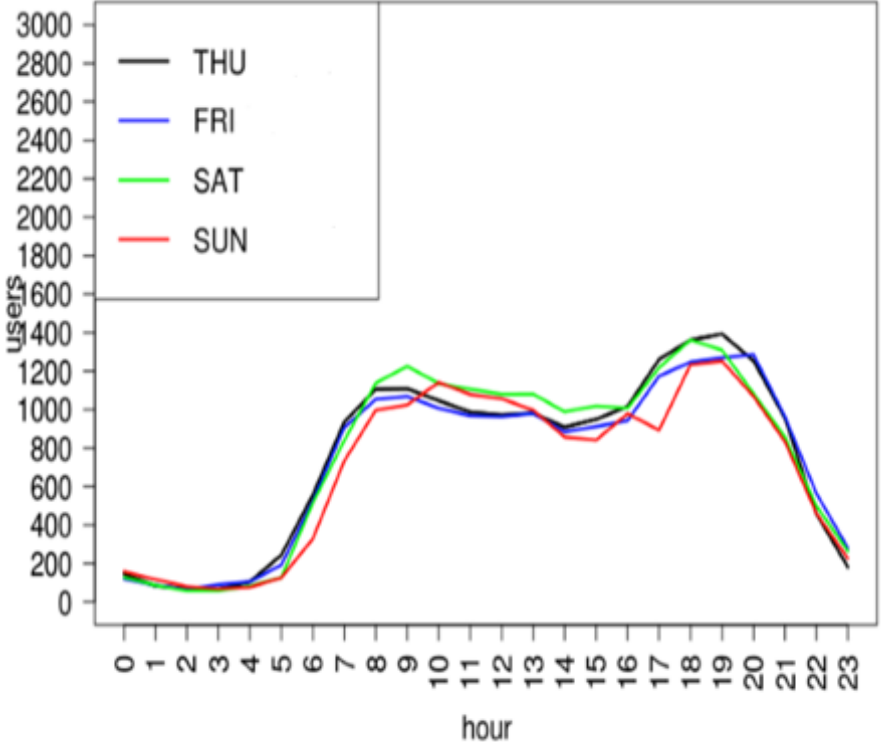
Understanding land use characteristics

- People leave digital traces when they use communication devices.
- Mobile communication patterns at different locations can be leveraged to classify them into land-use categories.

The diurnal patterns of users connected to base stations revealed interesting patterns



Base station 1

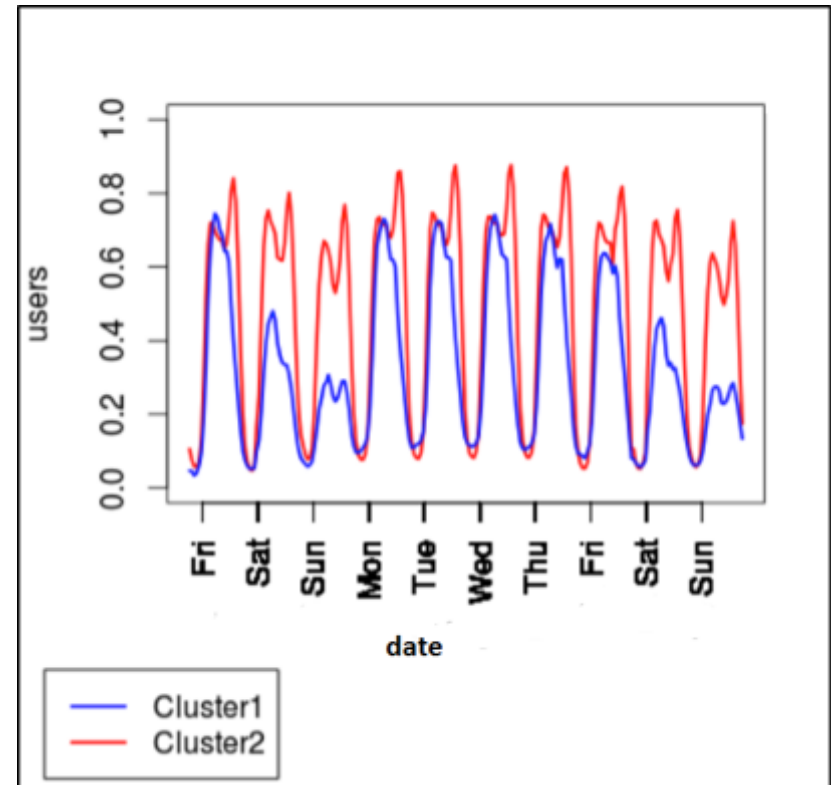
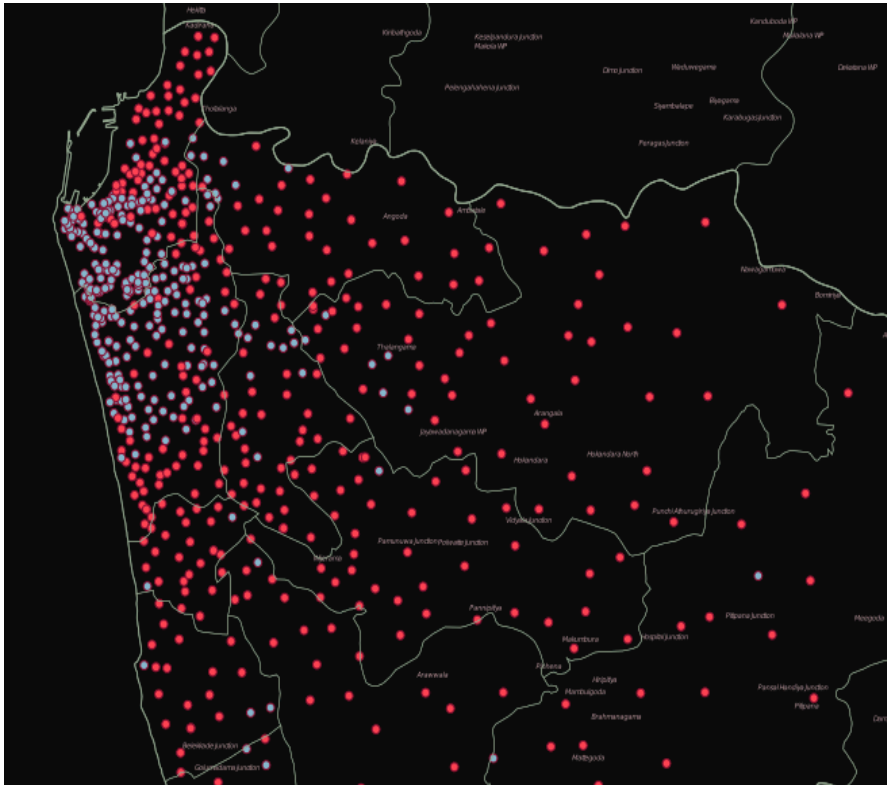


Base station 2

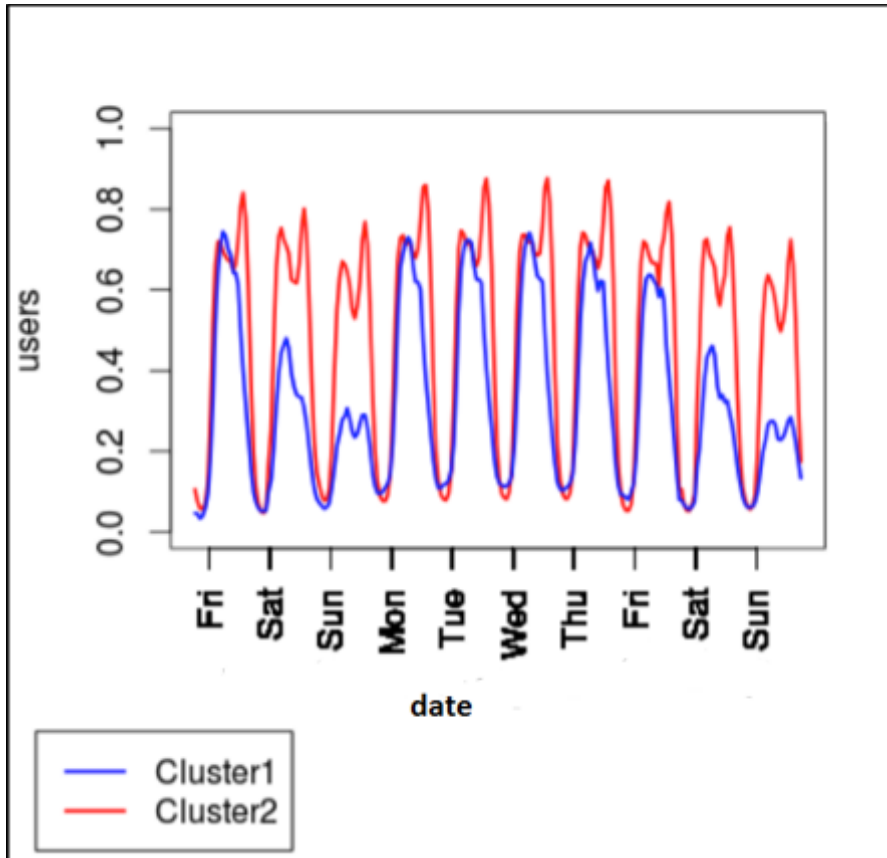
How do we leverage this insight?

- Time series of each base station is normalized to a (0-1) range
- Euclidean distance between two time series is used to cluster base stations in an unsupervised manner using k-means algorithm

Distribution of base stations in Colombo district



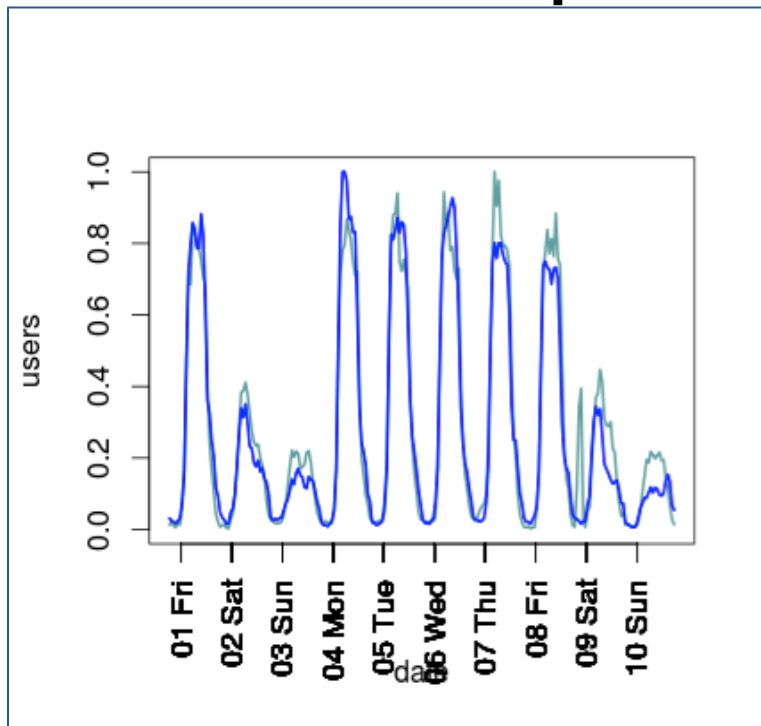
What does this reveal?



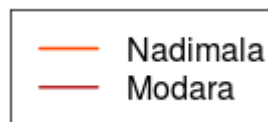
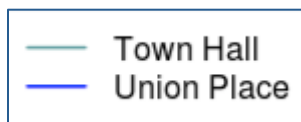
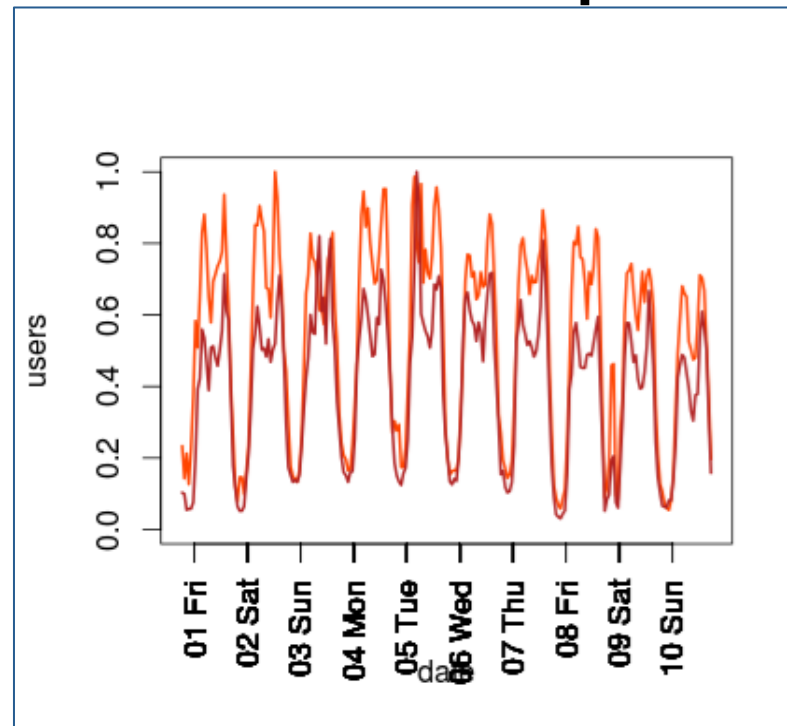
- Cluster 1 exhibits patterns consistent with a commercial area
- Cluster 2 exhibits patterns consistent with a less commercial and more residential characteristics (or possibly mixed)

A closer look at base stations in each cluster

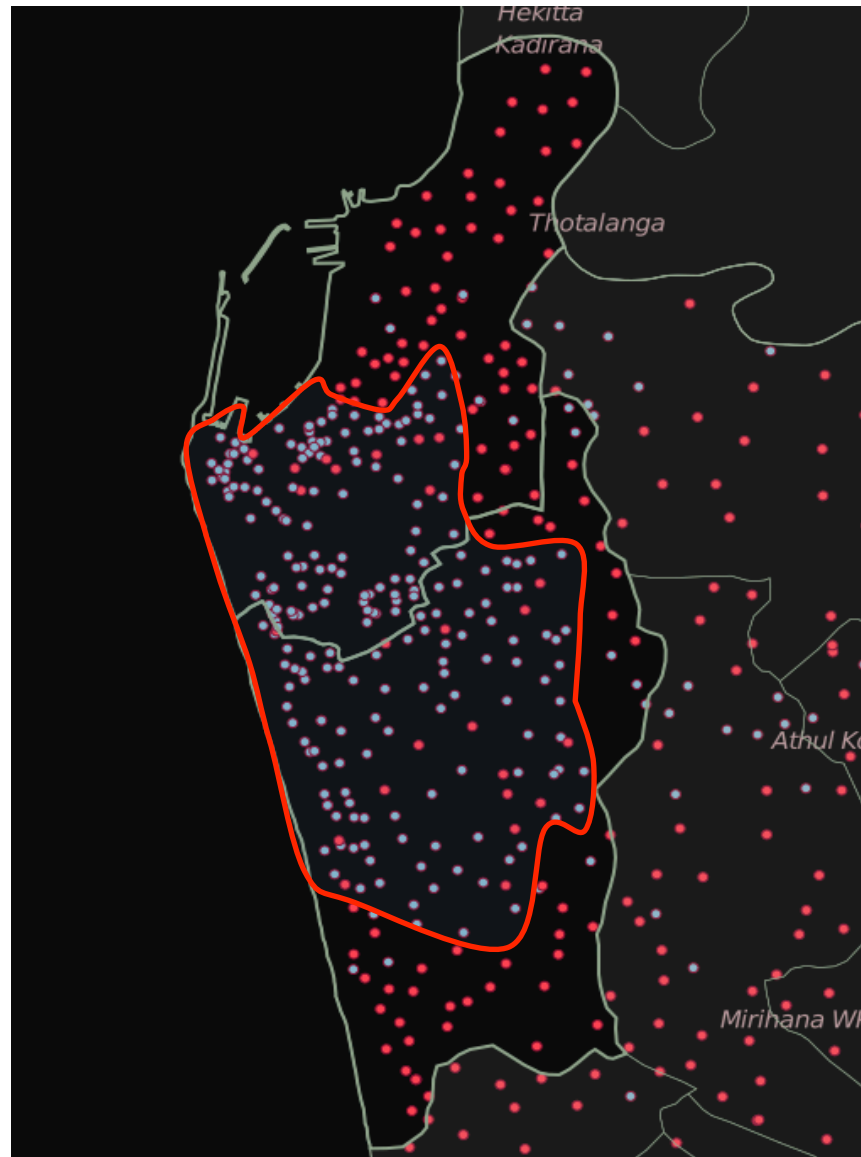
Cluster 1 examples



Cluster 2 examples



Our results show how the Central Business District (CBD) has expanded



Seethawaka Export Processing Zone (EPZ)



Photo ©Senanayaka Bandara - [Panoramio](#)

Seethawaka EPZ

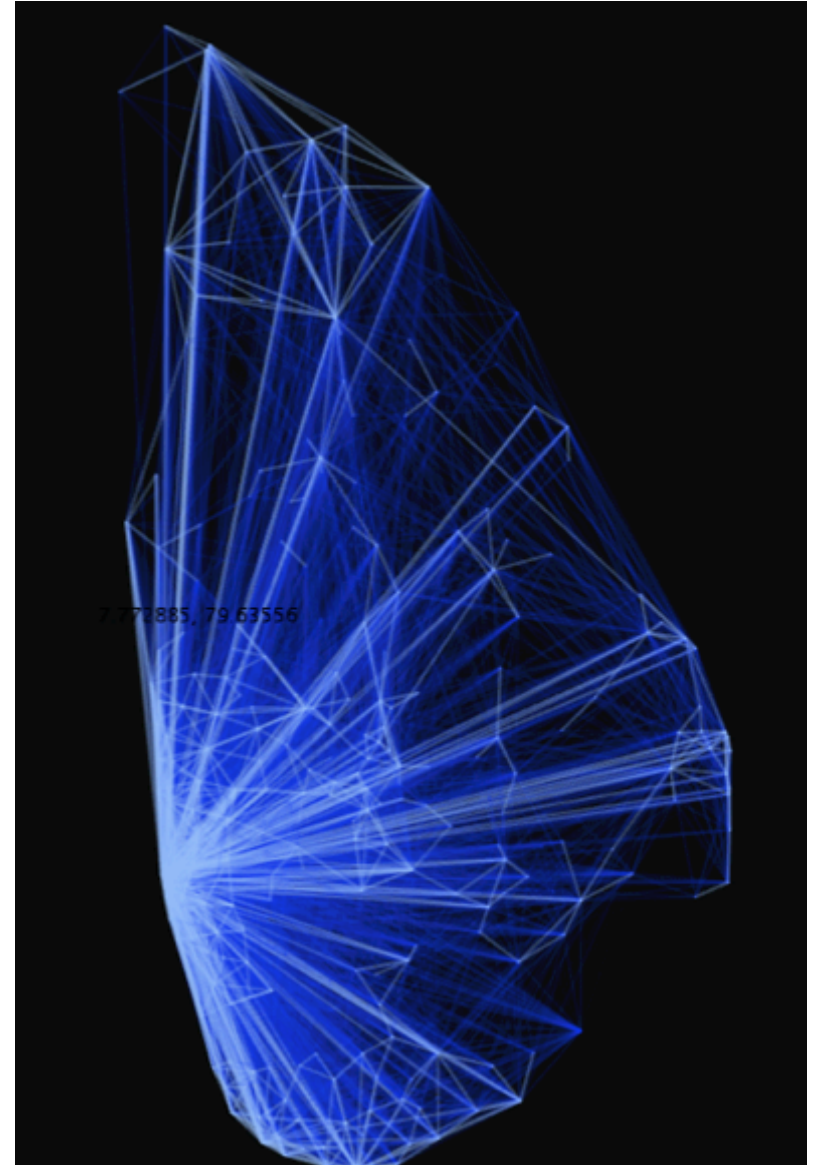
So how can mobile network big data help transportation and urban policy

- Can help us understand population mobility at a very high resolution (both temporally as well as spatially)
 - VLR data is very useful for understanding traffic congestion
- Can help us understand land-use characteristics
- Can provide insights in-between infrequent and costly surveys and land use census and in some cases even replace them

Can mobile network big data help us understand Sri Lanka's communities?

The geo-spatial distribution of Sri Lanka's social networks?

- Each link represents the raw number of outgoing and incoming calls between two DSDs
 - Divisional Secretariat Division (DSD) is a third level administrative division; 331 in total in LK

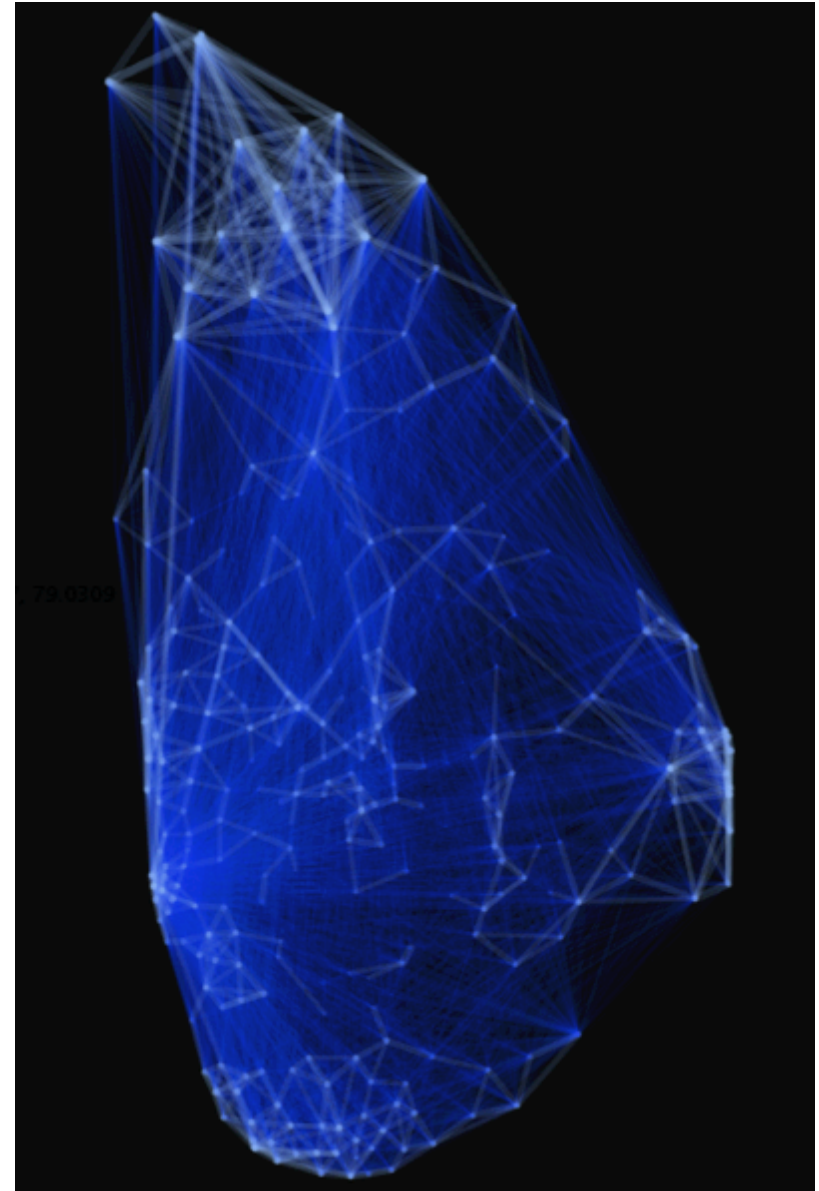


Low  High
No. of calls

A different picture emerges when call volume is normalized by population

$$\text{Normalized calls } (DSD_1, DSD_2) = \frac{\text{No. of calls } (DSD_1, DSD_2)}{\text{Population } (DSD_1) \times \text{Population } (DSD_2)}$$

- Strongly connected components are visible



Low  High
No. of calls 10

Identifying communities: methodology

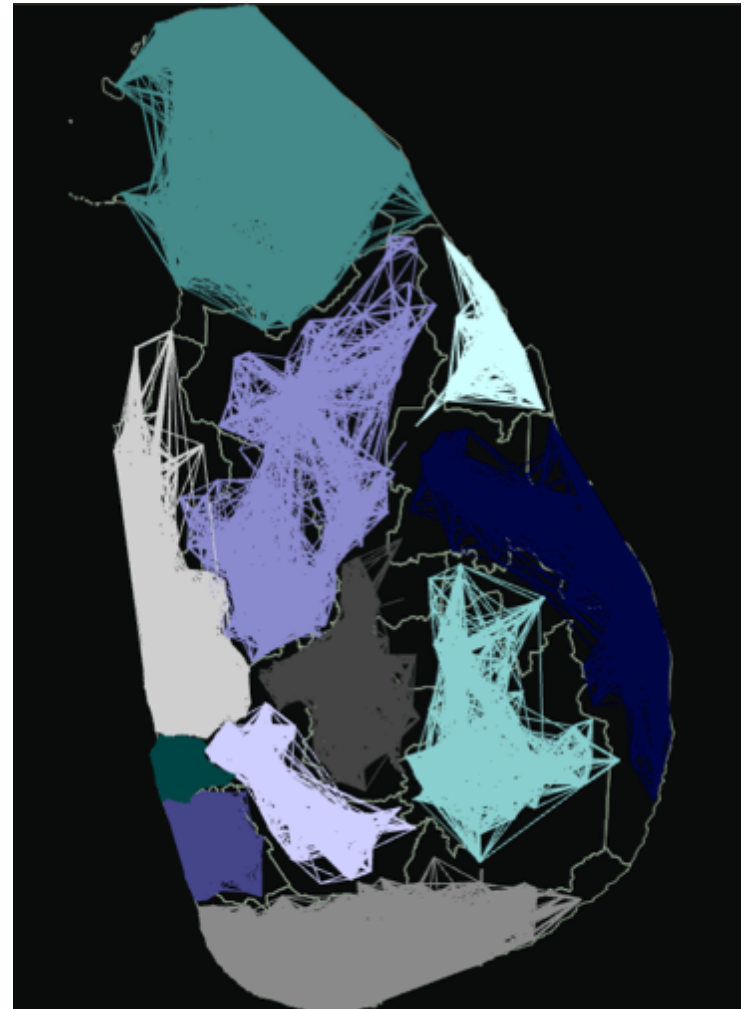
- The social network is segregated such that overlapping connections between communities are minimized.
- Strength of a community is determined by *modularity*
 - Modularity Q = (edges inside the community) –
(expected number of edges inside the community)

$$Q = \frac{1}{2m} \sum_{a,b} \left(A_{a,b} - \frac{k_a k_b}{2m} \right) \delta(c_a, c_b)$$

M. E. J.-Newman, Michele-Girvan, "Finding and evaluating community structure in networks", Physical Review E, APS, Vol. 69, No. 2, p. 1-16, 204.

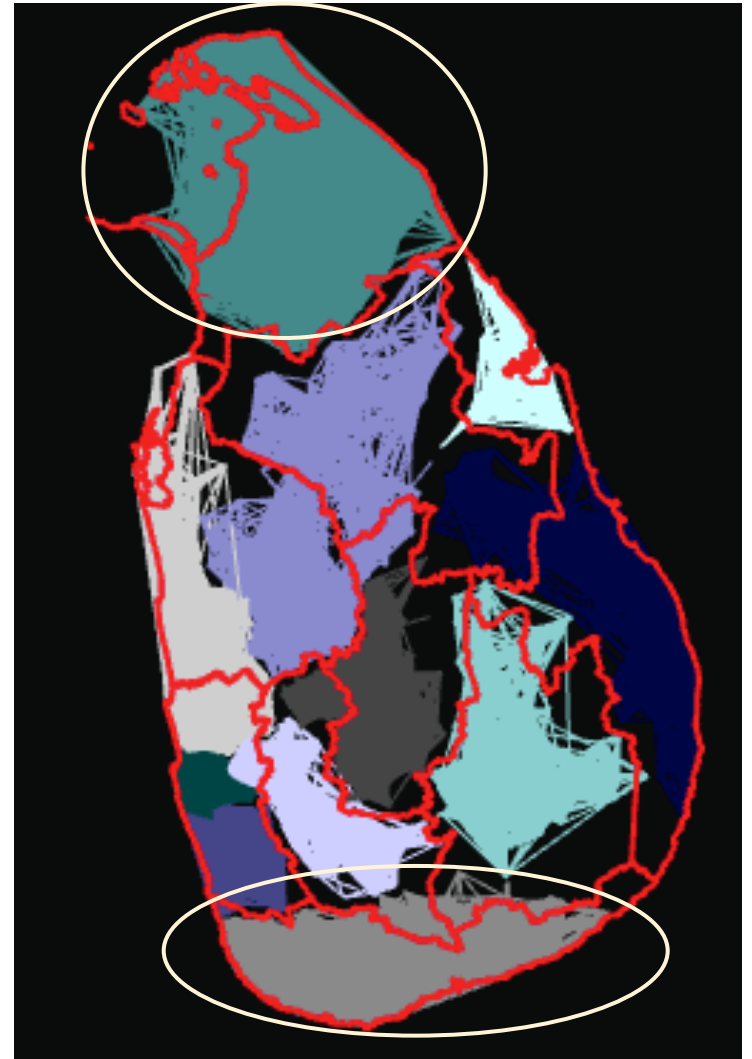
The resultant communities are centered around geographic neighborhoods

- For Sri Lanka, the optimal number of communities discovered by the algorithm was 11



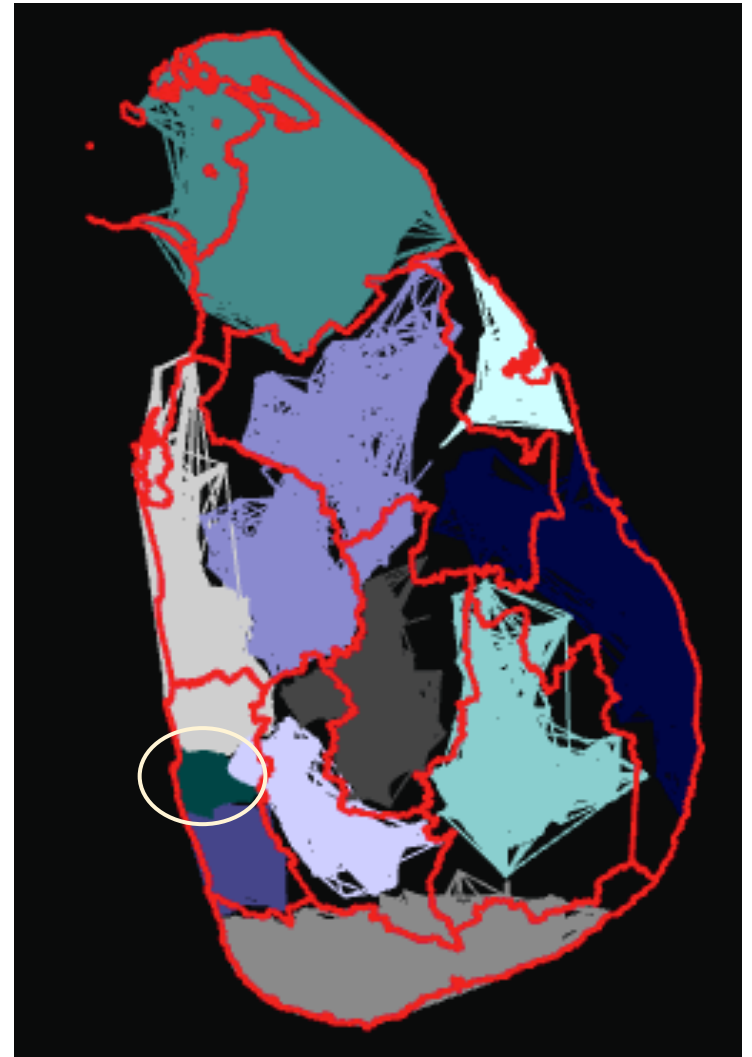
How much do these communities mesh with existing administrative boundaries?

- Southern and Northern provinces have the highest similarity to their respective provincial boundaries.



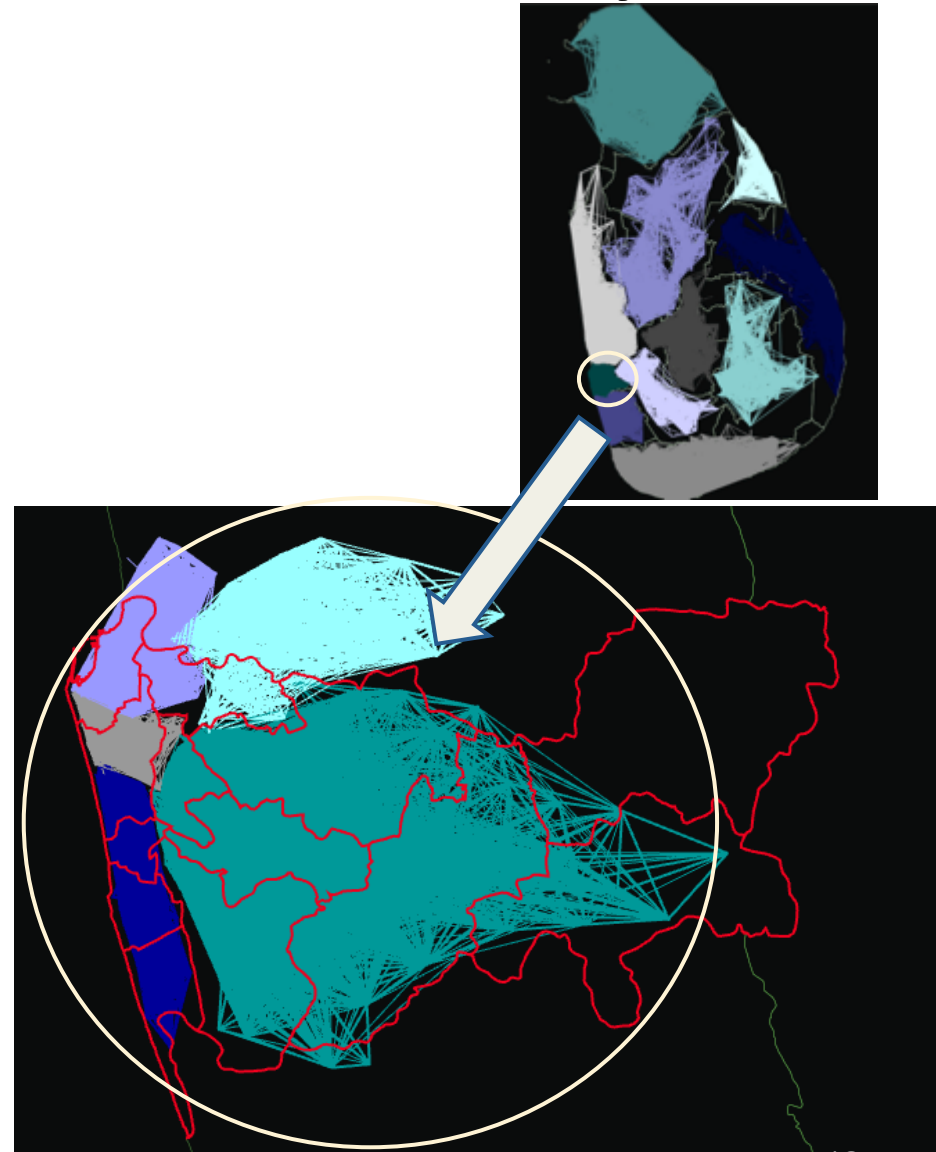
How much do these communities mesh with existing administrative boundaries (contd.)?

- Colombo district is clustered as a single community and Gampaha is merged with North Western Province



Zooming into a community

- Community detection is done within identified communities
- These communities are less related to DSD boundaries in Colombo district



Implications for public policy?

- Administrative boundaries based on history and geography may not reflect current community structures

Analytical challenges

Challenge	Solution(s)
Data is biased towards frequent users	<ul style="list-style-type: none">• Understand and adjust for selection bias
Data sparsity	<ul style="list-style-type: none">• Interpolation techniques• Probability based models
Different tower densities	<ul style="list-style-type: none">• Different scale of analyses depending on region
Validating results	<ul style="list-style-type: none">• Using other data sources e.g. data from Dept. of Census and Statistics, transportation survey data, etc.

In sum..

- Mobile network big data has many uses for developmental policy:
 - Helping us understand human mobility at a fine temporal and geo-spatial scale
 - Helping us understand social connectedness and community structures
- But it is not without analytical challenges
 - Incorporating other data (even ‘small data’) can improve the results

Thank you

- More information:
 - <http://lirneasia.net/projects/bd4d/>