## Big Data for Development in Sri Lanka

Sriganesh Lokanathan, Danaja Maldeniya

Colombo Big Data Meetup, Virtusa Auditorium 26 April 2016





This work was carried out with the aid of a grant from the International Development Research Centre, Canada and the Department for International Development UK.



## Our mission

<u>Catalyzing policy change through research to improve</u> <u>people's lives</u> in the emerging Asia Pacific by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology.



## Where we work

Big data work only in Sri Lanka in 2012-16 Extending to Bangladesh 2016 onwards



## Big data

- An all-encompassing term for any collection of data sets so large or complex that it becomes difficult to process using traditional data processing applications.
- Challenges include: analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations.
- Examples:
  - 100 million Call Detail Records per day generated by Sri Lanka mobile operators
  - 45 Terabytes of data from Hubble Telescope



## Why big data? Why now?

- Proximate causes
  - Increased "datafication": Very large sets of schema-less (unstructured, but processable) data now available
  - Advances in memory technology: No longer is it necessary to archive most data and work with small subset
  - Advances in software: MapReduce, Hadoop



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
- Commercial transactions (transaction-generated data)
  - E.g., Stock exchange data, bank transactions, credit card records, supermarket transactions connected by loyalty card number
- Sensors and tracking devices
  - E.g., road and traffic sensors, climate sensors, equipment & infrastructure sensors, mobile phones communicating with base stations, satellite/ GPS devices
- Online activities/ social media

– E.g., online search activity, online page views, blogs/ FB/ twitter

## Currently only mobile network big data has broad population coverage

	Mobile SIMs/100	Internet users/100	Facebook users/100
Myanmar	50	2	12
Banglade sh	76	10	9
Pakistan	73	14	11
India	73	18	9
Sri Lanka	107	26	16
Philippine s	112	40	41
Indonesia	125	17	25
Thailand	Source: ITU Measuring 143	Information Society 2015: Ea 35	cebook advantage portal 49

Pro-poor Pro-market

#### 



## Data used in the research

- Multiple mobile operators in Sri Lanka have provided four different types of meta-data
  - Call Detail Records (CDRs)
    - Records of calls
    - SMS
    - Internet access
  - Airtime recharge records
- Data sets do not include any Personally Identifiable Information
  - All phone numbers are pseudonymized
  - LIRNE*asia* does not maintain any mappings of identifiers to original phone numbers
- Cover 50-60% of users; very high coverage in Western (where Colombo the capital city in located) & Northern (most affected by civil conflict) Provinces, based on correlation with census data



#### Work performed by collaborative inter-disciplinary teams

#### • LIRNEasia

- Danaja Maldeniya
- Dedunu Dhananjaya
- Isuru Jayasooriya
- Kaushalya Madhawa (moved on to Tokyo Institute of Technology)
- Madhushi Bandara
- Nisansa de Silva (moved on to U of Oregon)
- Prof. Rohan Samarajiva
- Sriganesh Lokanathan
- University of Moratuwa
  - Dr. Amal Shehan Perera
    - Data Mining
  - Chatura de Silva
    - Urban Planning
  - Prof. Amal Kumarage
    - Transport
  - Undergraduates working on projects

- LIRNEasia/ MIT
  - Gabriel Kreindler (Economics)
  - Yuhei Miyauchi (Economics)
- Other US Universities
  - Prof. Joshua Blumenstock (U Washington, School of Information)
    - Data Science
  - Saad Gulzar (NYU)
    - Political Science
- Advisory Group:
  - Prof. Louiqa Rashid (U of Maryland)
  - Dr Prabir Sen (former Chief Data Scientist, IDA Singapore)
  - Dr Ruwan Weerasinghe (U of Colombo)
  - Prof Ryosuke Shibasaki (U of Tokyo)
  - Dr. Srinath Perera (WSO2)

## The technology used

- We built our own internal Apache Hadoop cluster:
  - 2 Master Nodes & 8 Slave Nodes
  - Total of 30 TB disk space with a replication factor of 2
- Distributed processing frameworks:
  - Apache Pig
  - Apache Hive
  - Apache Giraph
- Tools & libraries:
  - Java & Python
  - R
  - Processing
  - QGIS



## Apache Hadoop

- An open source distributed file storage and processing system based on Google's MapReduce and Google file System
- Map-Reduce the processing paradigm of Hadoop is a key-value based approach
  - Map: split data in to buckets based on a specified key value column
  - Reduce: perform specified operations on each bucket. Ex: generate summary statistics
- Native programming interface is Java with support for other languages like Python over the streaming API (less efficient).



## Apache Pig

- Apache Pig is a tool that provides a SQL like programming interface on top of Hadoop (Pig Latin)
- Scripts in Pig Latin are converted in to a series of map reduce jobs by the compiler
- Very useful for aggregation related tasks on large datasets
- More complex operations that are not natively supported can be easily added with custom functions written in java
- Not the first choice for implementing complex models/algorithms



## R

- A programming language meant for statistical computing and visualization
- Very popular among data scientists and researchers
- Extensive libraries for statistical analysis, machine learning, network analysis and visualization
- The eco-system provides a high level of support for high quality reproducible research/analytics
  - R Studio (a complete IDE)
  - R markdown and Knitr
  - Shiny applications
  - Plotting tools : ggplot, plotly etc.



## QGIS

- A open source geographical information system
- Provides extensive spatial analytics and visualization capabilities
- Used for creating map layers with different analytical results for visual inspection/analysis

## Processing

- General purpose visualization language/library built on top of Java
- Particularly useful for complex geospatial visualizations (E.g. animating traces of vehicles from the New York (taxidata)

## **Basic analytical flow**



## The rest of the presentation

- Understanding land use characteristics
- Measuring urban economic activity
- Understanding Sri Lankan communities
- Other interesting applications
- Analytical challenges



#### • Understanding land use characteristics

- Measuring urban economic activity
- Understanding Sri Lankan communities
- Other interesting applications
- Analytical challenges



## Hourly loading of base stations reveals distinct patterns



 We can use this insight to group base stations into different groups, using unsupervised machine learning



### Methodology

- The time series of users connected at a base station contains variations, that can be grouped by similar characteristics
- A month of data is collapsed into an indicative week (Sunday to Saturday), with the time series normalized by the z-score
- Principal Component Analysis(PCA) is used to identify the discriminant patterns from noisy time series data



- Each base station's pattern is filtered into 15 principal components (covering 95% of the data for that base station)
- Using the 15 principal components, we cluster all the base stations into 3 clusters in an unsupervised manner using k-means algorithm



## Three spatial clusters in Colombo District



- Cluster-1 exhibits patterns consistent with commercial area
- Cluster-3 exhibits patterns consistent with residential area
- Cluster-2 exhibits patterns more consistent with mixeduse



#### Our results show Central Business District (CBD) in Colombo city has expanded



## Small area in NE corner of Colombo District classified as belonging to Cluster 1?





## We use silhouette coefficients to understand the quality of the clustering

• Silhouette coefficient indicates quality of clustering

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

a(i) - average distance of i with all other
data within the same cluster
b(i) - average distance of i with all other
data within the neighboring cluster

Based on the s-values, Cluster
 3 is the least coherent amongst
 the three

Cluster	Avg. Silhouette Coefficient
1 – Commercial	0.46
2 – Residential	0.36
3 – Mixed-use	0.22



#### Internal variations in mixed use regions: More commercial or more residential?

• To evaluate the relative closeness to the other two clusters, we define extent of commercialization as:



Blue dots: more residential than commercial residential

Red dots: more commercial than

## Next we analyzed the variation of commercial and residential model with a more formal model

- We established average temporal signatures for (almost) purely commercial (C) and residential behavior (R) using a set of known locations (seed data)
- A linear model was fit to the temporal signature of each base station in the country

$$S_i = \alpha C + (1 - \alpha)R + \varepsilon$$

Where,

- $S_i$  Temporal signature of base station i
- C Commercial signature
- *R* Residential signature
- $\varepsilon$  Error



### Commercial to residential spectrum





Highly commercial

## Implications for urban policy

- Almost real-time monitoring of urban land use
  - We are currently working on understanding finer temporal variations in zone characteristics (especially the mixed-use areas)
- Can complement infrequent surveys & align master plan to reality
- LIRNE*asia* is working to unpack the identified categories further, e.g.,
  - Entertainment zones that show evening activity



- Understanding land use characteristics
- Measuring urban economic activity
- Understanding Sri Lankan communities
- Other interesting applications
- Analytical challenges



# What does mobility tell us about economic activity?

Economic activity = (number of workers) × (productivity per worker)

•We assume more productive regions are more attractive destinations

•Commuting patterns emerge from the trade-off between attractiveness of a workplace and the cost of getting there



## Example of commuting flows from one origin





## Theoretical model outline

Agent  $\omega$  at residential location *i* chooses work location *j* offering wage  $w_j$  and at distance  $d_{ij}$ , and  $\omega$  has effective income at *j* 

$$y_{ij\omega} = \frac{w_j z_{ij\omega}}{d_{ij}}$$

where  $z_{ij\omega}$  is iid Fréchet-distributed random productivity shock. Commuting flow probabilities:

$$\pi_{ij} = \frac{\left(w_j/d_{ij}\right)^{\epsilon}}{\sum_s (w_s/d_{is})^{\epsilon}}$$

We estimate origin-constrained gravity model:

$$\log(\pi_{ij}) = \psi_j + \epsilon \log(d_{ij}) - \mu_i + \varepsilon_{ij}$$



## Economic activity/km<sup>2</sup>



### Model validation using nightlight data from satellites Nightlights Mean income



# Incorporating other data can give further insights

#### Household data: Census/HIES/LFS Industrial data: ASI, Industrial Census

	Nightlights	Household data	Industrial Data
Geographic variation			
Time variation	yearly	quarterly/ 2-3yrs/decade	yearly/decade
Relevant variables		Education, (un)employment, skill levels	Employment, capital intensity
Ideal for:		Improving Measure	Improving & Validation

Benefit of an improved framework for modeling economic activity

- Increase the coverage of existing surveys (both temporal and geographic)
  - By calibrating with household, industry census and survey data, when available
  - Then, mobile data can be used to predict/extrapolate for time periods and regions without survey data
- Can capture informal economic activity
  - Other research suggests informal economy is almost 30% of GDP in Sri Lanka



- Understanding land use characteristics
- Measuring urban economic activity
- Understanding Sri Lankan communities
- Other interesting applications
- Analytical Challenges



Prima facie, Colombo city (Colombo & Thimbirigasyaya

DSDs) seems to be the center Sri Lanka's social network

- 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



No. of calls



## A different picture emerges when call volume is normalized by population

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

• Strongly connected regional networks become visible





## 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)

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

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.



### Sri Lanka is made up of 11 communities





## How do communities match existing administrative divisions?





The 9 provinces

# With some exceptions, boundaries of communities differ from existing administrative divisions

- Northern (1), Uva (10) and Southern (11) communities most similar to existing provincial boundaries; but 11 takes Embilipitiya and Kataragama
- Colombo district is clustered as a single community (7)
- Gampaha merges with coastal belt of North Western Province (2) and Kalutara (8) is its own community
  - What does this mean for Western Province Megapolis?
- Batticaloa & Ampara districts of the Eastern Province merge with the Polonnaruwa district of North Central Province to form its own distinct community (6)
  - Possibly reflective of economic linkages since this is the rice belt of Sri Lanka
  - Does economics override ethnicity?







#### More differences appear when we zoom in further



- The littoral regions form their own distinct subcommunities
- The northern part of Colombo city forms a community with Wattala, across the Kelani river
- In general, rivers no longer form natural boundaries of communities



- Understanding land use characteristics
- Measuring urban economic activity
- Understanding Sri Lankan communities
- Other interesting applications
- Analytical Challenges



## Other ongoing research

- Modeling infectious disease propagation (Dengue) based on human mobility from CDR
- Measuring the impact of a transport shock (Opening the E03 expressway)
- Traffic analysis using CCTV footage
- Enhancing land use predictions with social media data (Eg. Foursquare) and satellite imagery
- Modeling interactions between different land
- Developing socio-economic indices and poverty mapping using CDR, satellite imagery, census data, etc.



- Understanding land use characteristics
- Measuring urban economic activity
- Understanding Sri Lankan communities
- Other interesting applications
- Analytical Challenges



# Addressing analytical challenges

Challenge	Solution(s)
Data is biased towards frequent users	<ul> <li>Understand and adjust for selection bias</li> </ul>
Data sparsity	<ul><li>Interpolation techniques</li><li>Probability based models</li></ul>
Different tower densities	<ul> <li>Different scale of analyses depending on region</li> </ul>
Validating results	<ul> <li>Using other data sources, e.g., data from Dept. of Census and Statistics, transportation survey data</li> </ul>

#### Selected Publications & Reports

- Lokanathan, S., Kreindler, G., de Silva, N. D., Miyauchi, Y., Dhananjaya, D., & Samarajiva, R. (forthcoming). Using Mobile Network Big Data for Informing Transportation and Urban Planning in Colombo. *Information Technologies & International Development*
- Samarajiva, R., Lokanathan, S., Madhawa, K., Kriendler, G., & Maldeniya, D. (2015). Big data to improve urban planning. *Economic and Political Weekly*, Vol L. No. 22, May 30
- Maldeniya, D., Lokanathan, S., & Kumarage, A. (2015). Origin-Destination matrix estimation for Sri Lanka using mobile network big data. 13th International Conference on Social Implications of Computers in Developing Countries. Colombo
- Kreindler, G. & Miyauchi, Y. (2015). Commuting and Productivity: Quantifying Urban Economic Activity using Cell Phone Data. LIRNEasia
- Lokanathan, S & Gunaratne, R. L. (2015). Mobile Network Big Data for Development: Demystifying the Uses and Challenges. *Communications & Strategies*.
- Lokanathan, S. (2014). The role of big data for ICT monitoring and for development. In *Measuring the Information Society 2014*. International Telecommunication Union.

#### More information:

http://lirneasia.net/projects/bd4d/

