Towards understanding “Spend Potential” Flows from Mobile Money for Locating Retailers

Rui Cao,
Dr. Gavin Smith & Dr. James Goulding

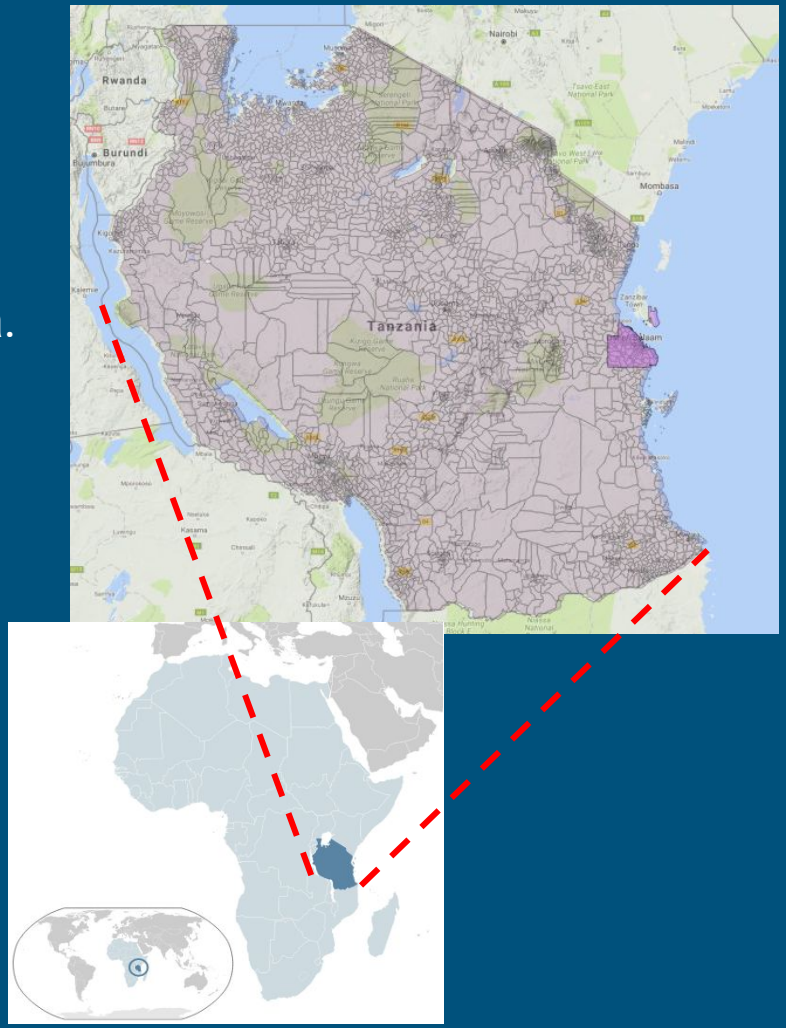
Background

- Survey of customer spending behaviour is very expensive, governments and companies have very limited related knowledge, especially for developing countries
- With the development of mobile financial payment system in developing countries, there are opportunities to capture real-world money flows based on generated data from them
- Have a deeper understanding of economic activities and behaviour in developing countries, thus in turn help the development of economy

Study Area

Dar es Salaam, a coastal city in Tanzania, Africa.

- Population: 4,364,541 (2012 census)
- Area: 1,590.5 km^2
- 1 GBP = 2800 TZS (Tanzanian shilling)
- Largest city in Tanzania and east Africa by population
- Tanzania's leading financial centre



https://en.wikipedia.org/wiki/Dar_es_Salaam

Study Data

Mobile phone money transaction records from a big mobile operator:

- Time span: 01/01/2014 - 12/31/2014
- 10% sample (over 700,000 users)
- Key fields:
  - Sender/receiver/customer id
  - Transfer amount
  - Transfer date/time
  - Transfer status
  - Service type
  - Sender/receiver category

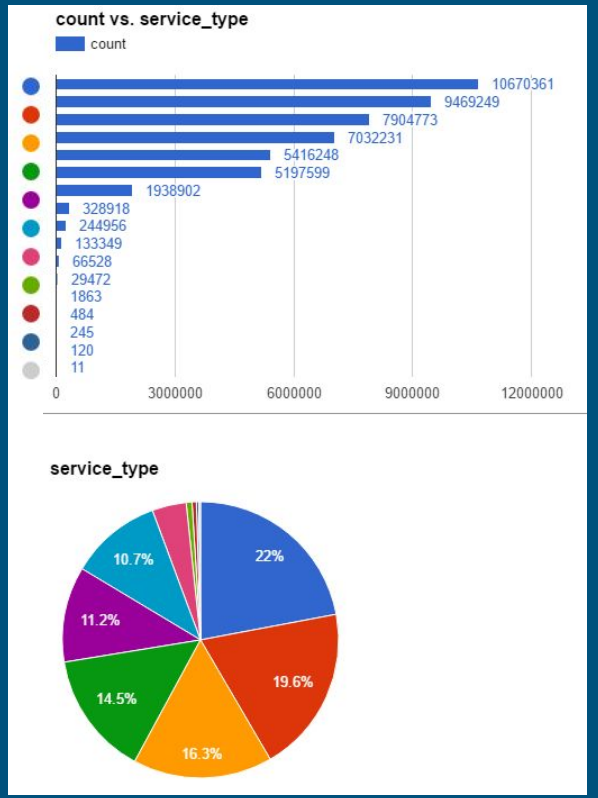
Aims & Objectives Overview

- Data exploration and preprocessing
- **User classification (distinguish sellers and non-sellers)**
- Augmentation of existing OD matrix & evaluation of potential retail locations

Data Exploration & Preprocessing

Service types:

- Money-transactional operations
  - Percentage: 65.80% (⅔)
- Non-money-transactional operations
  - Eg. Recharging, balance enquiry
  - Percentage: 33.99% (⅓)

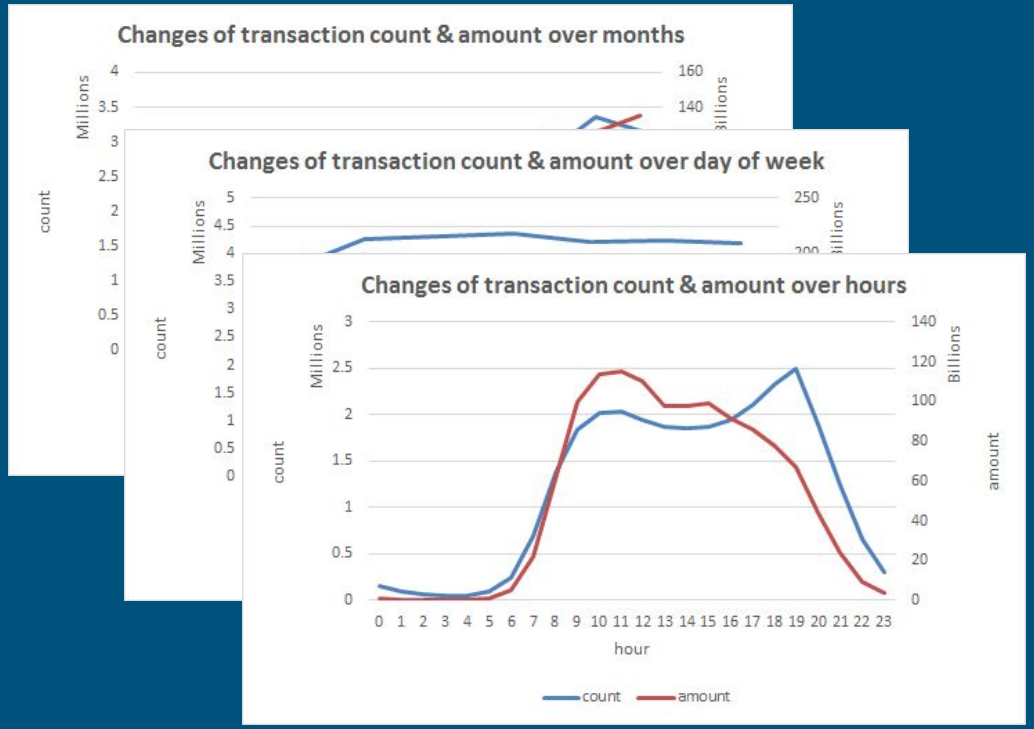


Data Exploration & Preprocessing

- Data preprocessing results overview:

	number	percentage
original	48,435,309	100%
unsuccessful	4,272,339	8.82%
zero-amount	1,340,588	2.77%
remaining	42,823,343	88.41%
selected	29,289,097	60.47%

Data Exploration & Preprocessing



- Changes over months in a year
  - Gradually increase with small fluctuations
- Changes over day of week
  - Transaction money reached the lowest point on Sunday while transaction count varied slightly
- Changes over hours in a day
  - Morning peak of transaction money
  - Night peak of transaction count

User Classification

- Characterize users by money transaction behaviour
- Distinguish sellers and non-sellers by cluster analysis on basis of user transaction characteristics

Characterize Money Transaction Behaviour

- Total count
- Daily count
- Count in weekdays
- Count in working hours

- Total days
- Total weekdays
- Distribution of duration

- Total amount
- Single amount
- Daily amount

- #distinct-senders
- #distinct-receivers

Cluster Analysis

Two-step strategy:

- Split users into three parts by thresholds first to avoid outliers in user attribute values
  - Lower boundary: total count < 1 per week;
  - Upper boundary: receiving count > 10 per day;
  - Middle part: in the middle of lower and upper boundaries
- Cluster analysis to separate confusing users
  - PCA transform and normalisation for variables of characteristics
  - Two-step clustering method by SPSS

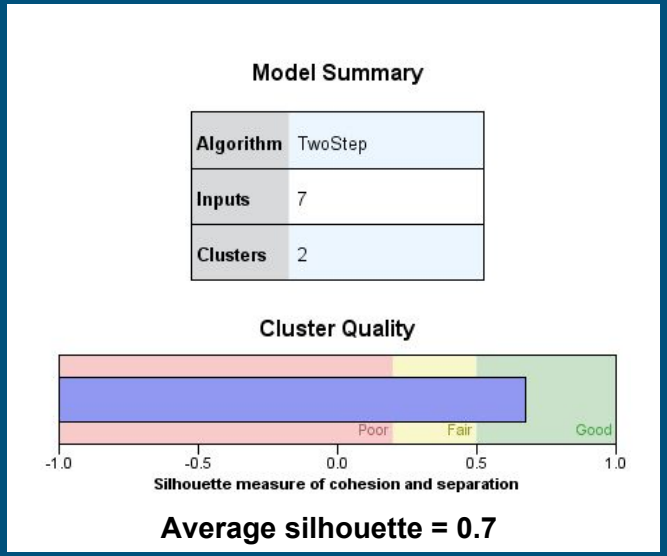
Cluster Analysis

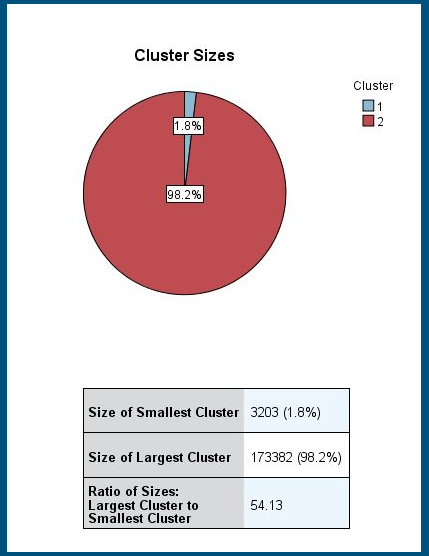
Results for splitting:

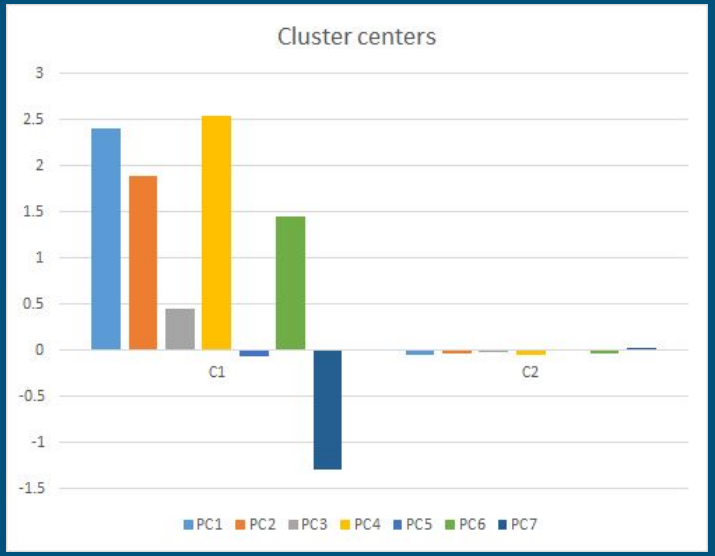
category	count	percentage of all	percentage of remaining
all	637295		
remaining (total day >= 3)	495331	77.72%	
lower (total count < 1 per week)	300701	47.18%	60.71%
upper (recv count > 10 per day)	530	00.08%	00.10%
for clustering	194100	30.46%	39.19%

Preliminary Results & Conclusions

- Clustering results overview:

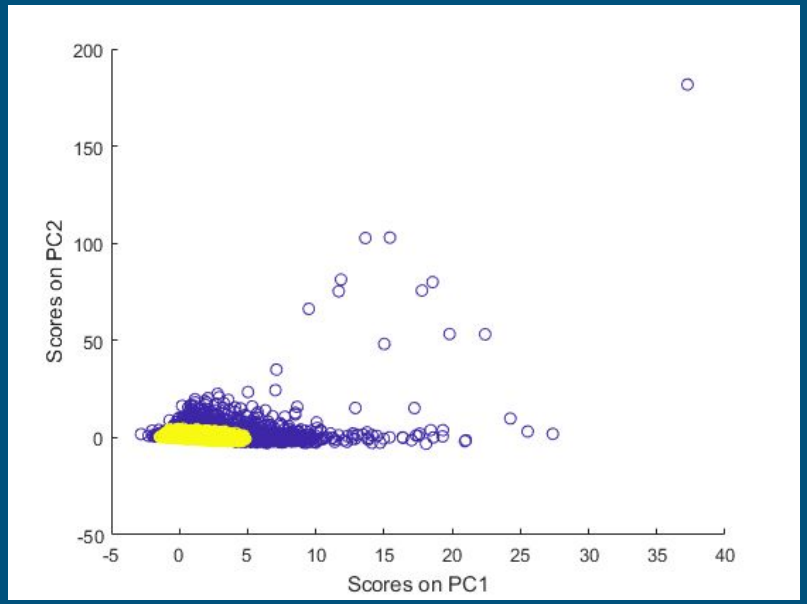


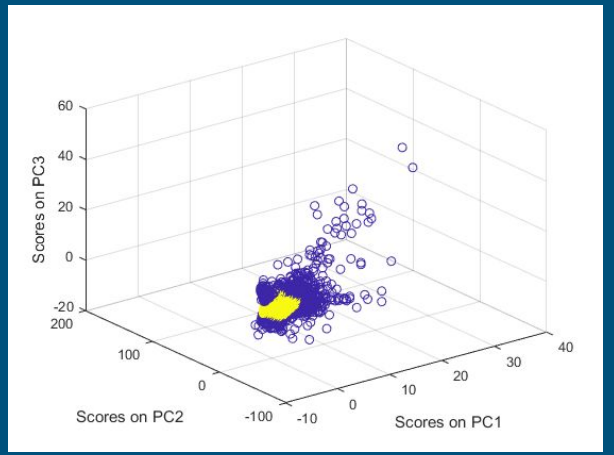




Preliminary Results & Conclusions

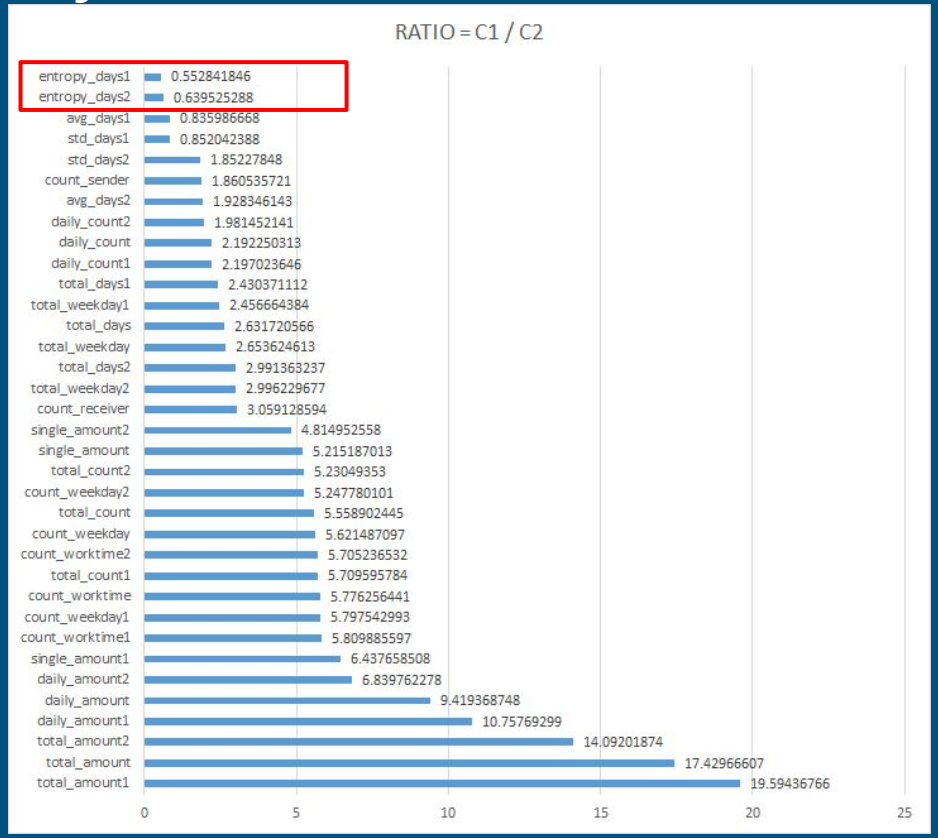
- Visualisation of primary components:





PC1: Count + Time + Interaction
PC2: Money

Preliminary Results & Conclusions

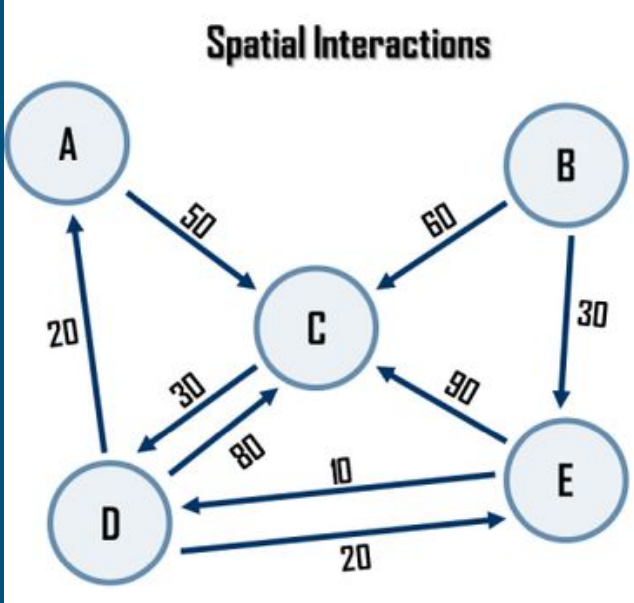


Preliminary Results & Conclusions

What we can learn from clustering results:

- Most users are non-sellers.
- Transaction record count, and transfer money amount of sellers are all significantly larger than non-sellers.
- Sellers interact with much more users.
- Transactions of sellers happened more frequently and more regularly.

Following Work & Expected Outcomes

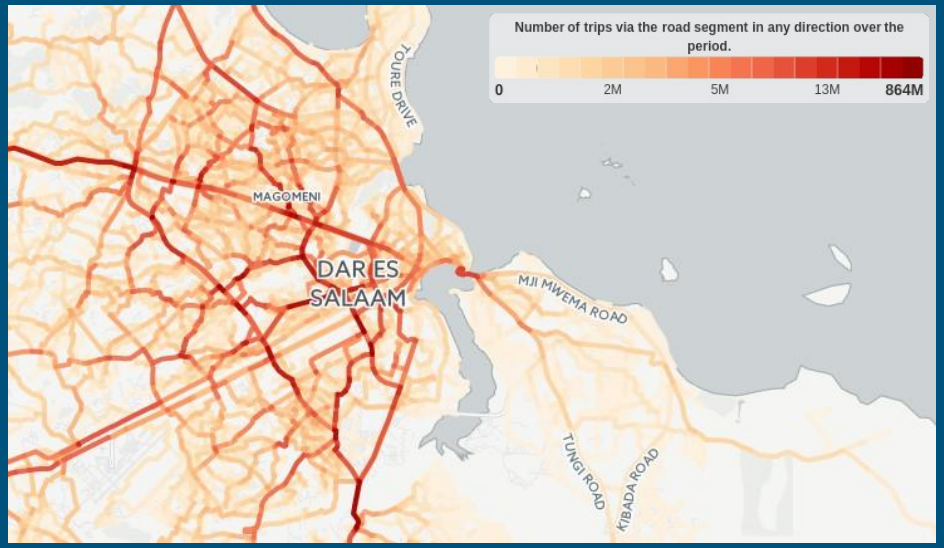


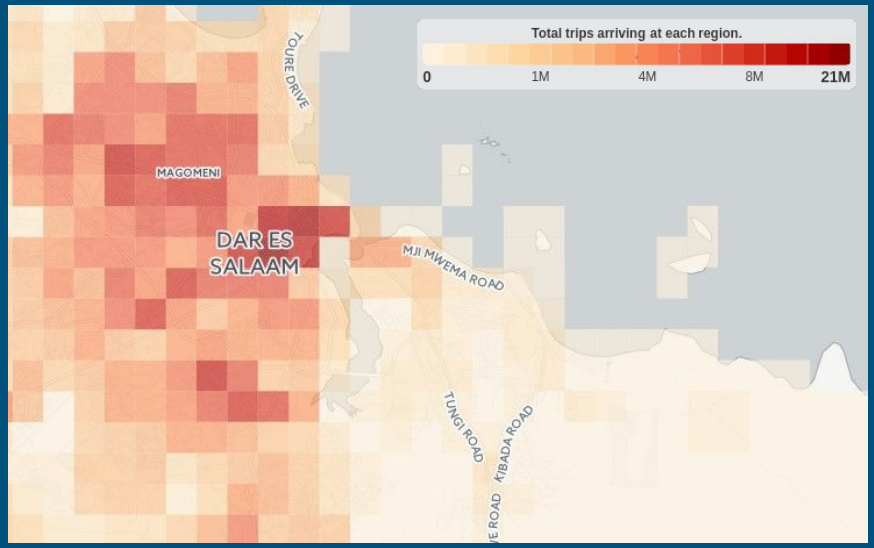
O/D Matrix

	A	B	C	D	E	Ti
A	0	0	50	0	0	50
B	0	0	60	0	30	90
C	0	0	0	30	0	30
D	20	0	80	0	20	120
E	0	0	90	10	0	100
Tj	20	0	280	40	50	**390**

Following Work & Expected Outcomes

Visualisation of OD matrix:





Ethical Issues

Ethical issues mainly surround the use of CDR and mobile money data.

Several measurements are taken to avoid unethical use:

- Research outputs are always aggregated
- Attempts of re-identification will not be considered
- Data storage, access, and usage are strictly monitored and encrypted

Reflections

- Real world data is dirty and full of "surprise" in terms of missing, messy and outlier data.
- Data preprocessing is very time-consuming but important in *Big Data* research.
- Data mining should be associated with expert knowledge.
- Expectation is always too good to be true, flexible working plan and patience are always needed.

Thank you!

Q & A