Big Data Recommendation Approaches for Healthcare

Samee U. Khan

Department of Electrical and Computer Engineering
North Dakota State University
Fargo, ND 58108-6050, USA
samee.khan@ndsu.edu

Outline

- Introduction – Personal & Topic
- Recommendation system models
- Big data recommendation system applications
- Case studies

Recommendation Systems and Big Data

Recommendation Systems

Introduced in 90s:

- Information filtering
- Personalization
- Recommend items/services

Perspectives

Customers' perspective

- Finding items of interest
- Narrow down choices
- Customizations
- Predict needs

Providers' perspective

- Understanding customers' behavior
- Increase sales
- Product promotion
- Trend analysis

amazon ebay

Route XL

FOURSQUARE

Big Data

Recent Web trends require tools and methodologies to efficiently manage the data for curation, processing, and storage

Challenges

- Storage
- Availability
- Reliability
- Computations
- Scalability

Big Data

Dimensions

- Volume
- Velocity
- Variety
- Veracity

Ever growing data e.g. 500 million tweets daily [1] and 600 TB daily on Facebook[2]

Time sensitive applications e.g. scrutinize fraud from millions of trade events

Structured (relational data), unstructured (text, audio, video, log files etc.). 80% unstructured[3]

Data authenticity and correctness

[1] "Internet Live Stats,"http://www.internetlivestats.com/twitter-statistics/, Accessed on April 10, 2018.

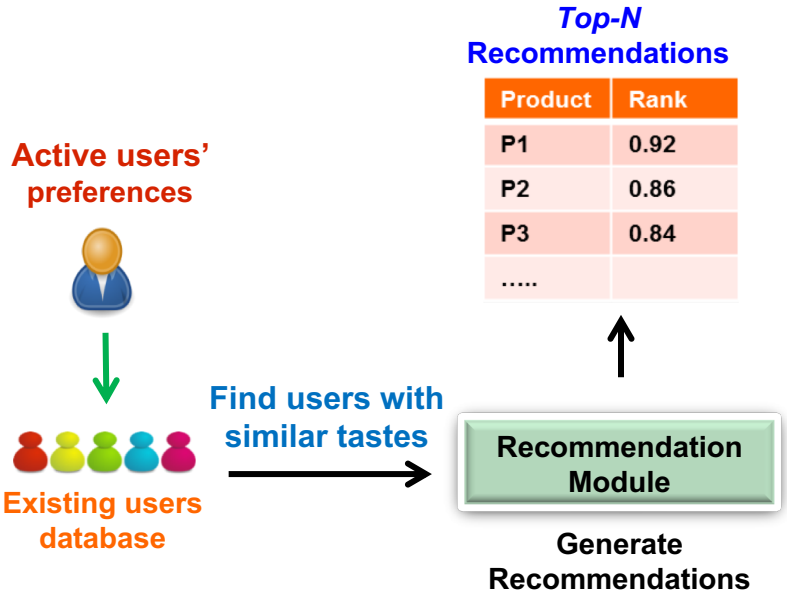
[2] "Fcode," https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/, Accessed on April 10, 2018.

[3] "Unstructured Data—A Growing Problem," https://www.waterfordtechnologies.com/unstructured-data-growing-problem/#more-10513, Accessed on April 10, 2018.

North Dakota State University NDSU

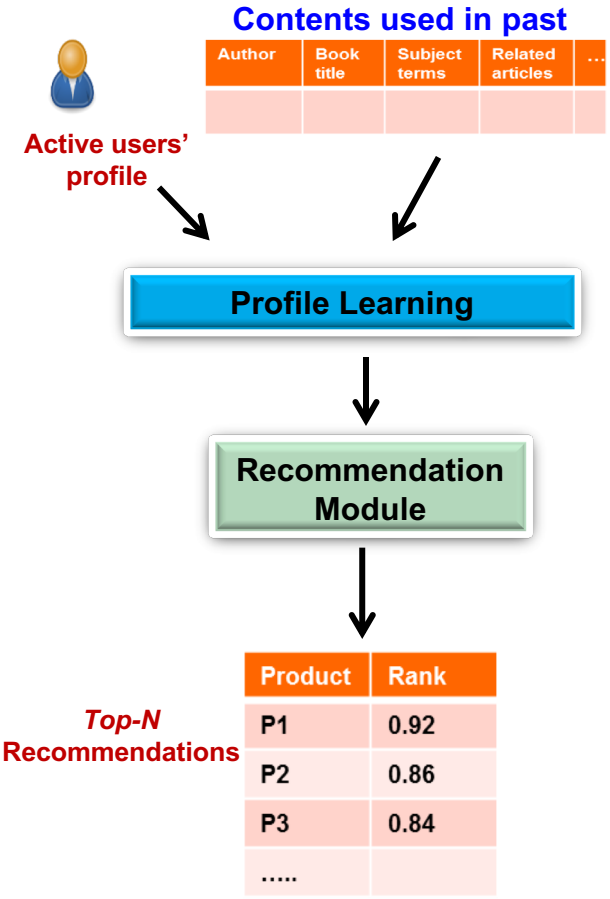
Recommendation System Models

- Collaborative Filtering
- Content Based Filtering
- Hybrid Filtering
- Collaborative Filtering
  - Information filtering through human behavior/user profiles
  - commonly employed in commercial recommender systems
    - Example: Amazon
- Issues with Collaborative Filtering
  - Cold Start
    - Requires enough users/items in the system
  - Sparsity
    - Occurs due to scarce data points
  - Long-tail/Popularity Bias
    - Recommendation of popular items only
  - Scalability
    - Occurs due to increase in users and items



Recommendation System Models

- Content Based Filtering
  - Recommendations based on the contents of items instead of users ratings or opinion
  - Requires no information about other users
  - Recommendations for users with unique tastes
  - No cold start and sparsity issues
  - Capable of recommending new or unpopular items
- Issues with Content Based Filtering
  - Requires meaningful encoding of content features
  - Inability to utilize judgement quality of other users
- Hybrid Filtering
  - Combination of collaborative and content based filtering
    - Popularity data
    - Contents
- Issues with Hybrid Filtering
  - Datasets interoperability



Big Data Recommendation Systems Applications

- Healthcare
  - Health expert recommendation from social media
  - Disease risk assessment (prediction)
  - Health insurance recommendation
- Route Recommendation Systems
  - Social venues
  - Large-scale evacuation

Case Study I: Personalized Healthcare Services[1]

- Increasing trends for finding online health information
  - Health related searches by 93 million Americans (Pew Internet & American Life Project)[2]
- Health information from online health communities
  - Exchange and share disease specific experiences
  - Psychological support from peers (example: patientslikeme[3])
- Increased expenditure of healthcare
  - U.S. healthcare expenses approx. 18.2% of the GDP till now (2018)[4]
- Key Contributions:
  - Disease risk assessment (NHANES 2009—2010 dataset[5])
  - Health expert recommendation from Twitter
    - 1,500,000,000+ Healthcare Tweets[6]
    - 30,000 provider profiles (MD, RN etc.)
    - 15,780 predefined topics
    - 16,283 health communities

[1]A. Abbas, M. Ali, M. U. S. Khan, S. U. Khan, "Personalized Healthcare Cloud Services for Disease Risk Assessment and Wellness Management using Social Media" *Pervasive and Mobile Computing,* vol. 28, pp. 81-96, 2016.

[2]"NBCNews," http://www.nbcnews.com/id/3077086/t/more-people-search-health-online/#.Ws4ss4hubIU, accessed on April 11, 2018.

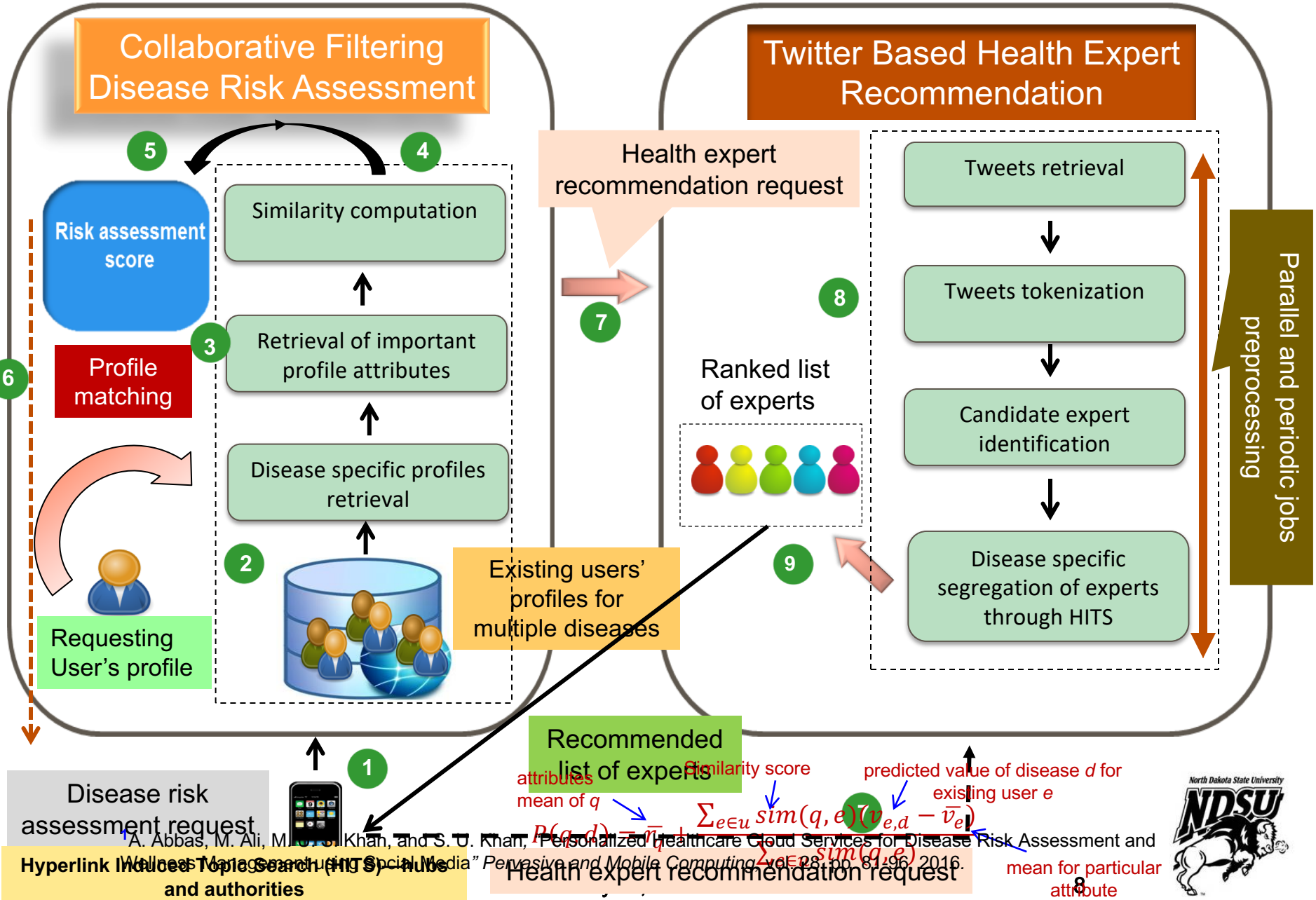
[3]"Patientslikeme",http://www.patientslikeme.com/, accessed on April 11, 2018.

[4] "Statista: The Statistical Portal," https://www.statista.com/statistics/184968/us-health-expenditure-as-percent-of-gdp-since-1960/, Accessed on April 11, 2018.

[5]"National Health and Nutrition Examination Survey," http://wwwn.cdc.gov/nchs/nhanes/search/nhanes09_10.aspx, accessed on September 29, 2014.

[6]"Healthcare Social Media Analytics," http://www.symplur.com/healthcare-social-media-analytics/, accessed on April 11, 2018.

Case Study I: Personalized Healthcare Services[1]



[1] A. Abbas, M. Ali, M. U. S. Khan, and S. U. Khan, "Personalized Healthcare Cloud Services for Disease Risk Assessment and Wellness Management using Social Media" *Pervasive and Mobile Computing*, vol. 28, pp. 81-96, 2016.

Case Study I: Personalized Healthcare Services[1]

User-keyword matrix

	K_1	K_2	K_3	K_4	K_5	K_6
U_1	5	1	2	2	5	1
U_2	-	3	2	8	2	-
U_3	3	1	2	4	6	-
U_4	4	-	-	-	-	11

Hub score

Iteration No.	U_1	U_2	U_3	U_4
1	0.281	0.218	0.255	0.234
38	0.275	0.249	0.288	0.196

Authority score

Iteration No.	K_1	K_2	K_3	K_4	K_5	K_6
1	0.197	0.060	0.067	0.235	0.246	0.191
39	0.190	0.065	0.068	0.258	0.254	0.163

Evaluation

TP=True Positive, FP=False Positive, TN=True Negative, FN=False Negative

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{F} - \text{measure} = 2.\frac{\text{precision.recall}}{\text{precision} + \text{recall}}$$

CFDRA Evaluation

CART— data partitioning to classify the presence of absence of disease

Logistic Regression— relationship between disease data attributes to determine outcomes

Naïve Bayes— probability for the presence or absence of disease

BF tree— best data split to determine the outcomes

BayesNet— DAG to represent the relationship between disease and symptoms

MLP—attributes provided at input layer to produce outp

Random Forest—creation of multiple trees

Rotation Forest — splitting of dataset and eval

SVM— hyperplane separates patients from non-pa

When it predicts YES, how often it is correct?

EUR Evaluation

RowSum —health related keyword count

[2]*Paul et al.* — topical authority identification (tweets, retweets, self-similarity)

[3]*Cheng et al.* — local experts in an area

When it's actually YES, how often does it predict YES?

Health related
from

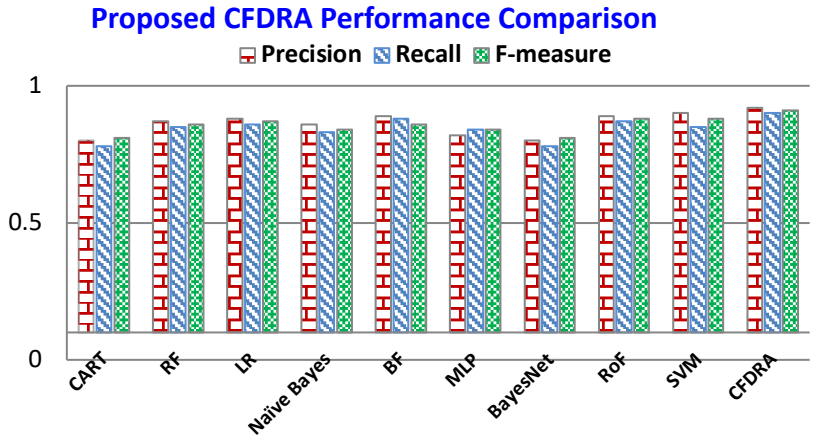
age", "gender", "ethnicity/race",
"weight", "diagnosed high blood sugar or pre-diabetes", "diabetes family history", "physical activity", "ever observed high blood pressure", "blood cholesterol", and "smoking, "ever diagnosed diabetes".

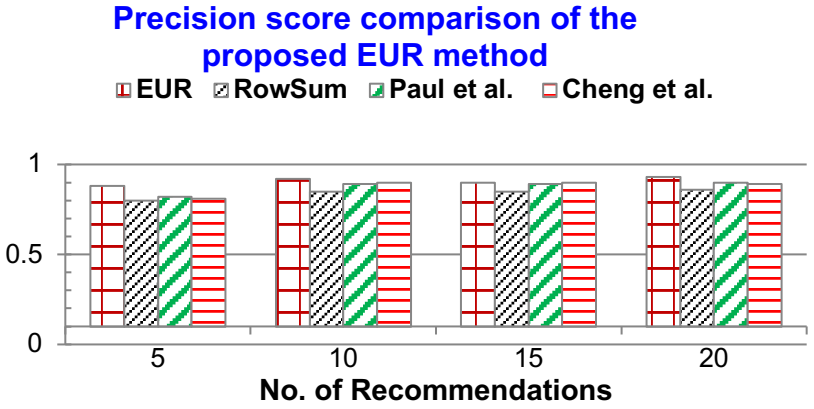
[1]A. Abbas, M. Ali, M. U. S. Khan, and S. U. Khan, "Personalized Healthcare Cloud Services for Disease Risk Assessment and Wellness Management using Social Media" *Pervasive and Mobile Computing,* vol. 28, pp. 81-98, 2016.

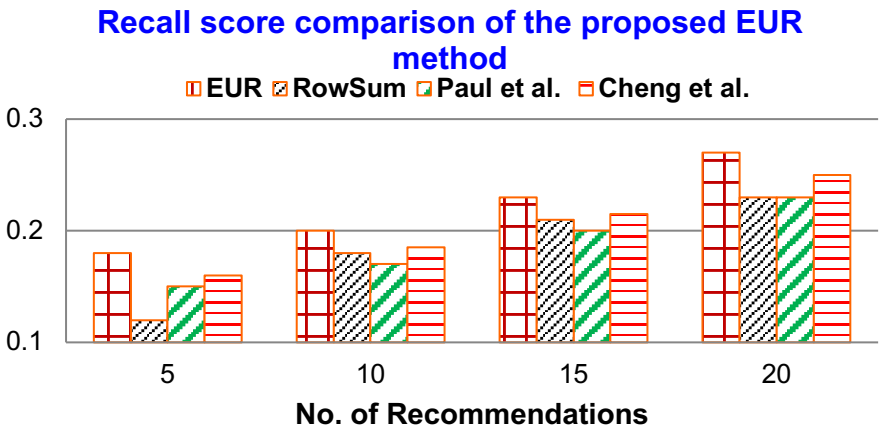
[2] A. Paul, and S. Counts, "Identifying topical authorities in microblogs," In *Proceedin Web search and data mining*, 2011, pp. 45-54.

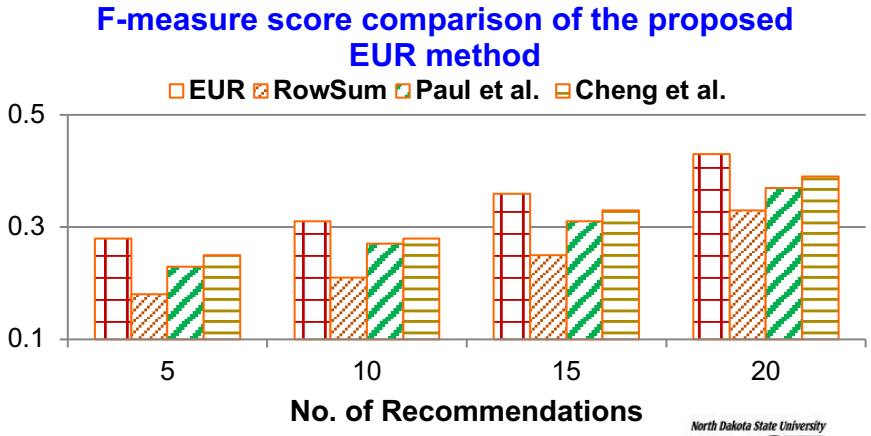
[3]Z. Cheng, J. Caverlee, H. Barthwal, and V. Bachani, "*Who is the Barbecue King of Experts on Twitter*," In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2011, pp. 335-344.

Case Study I: Experimental Results



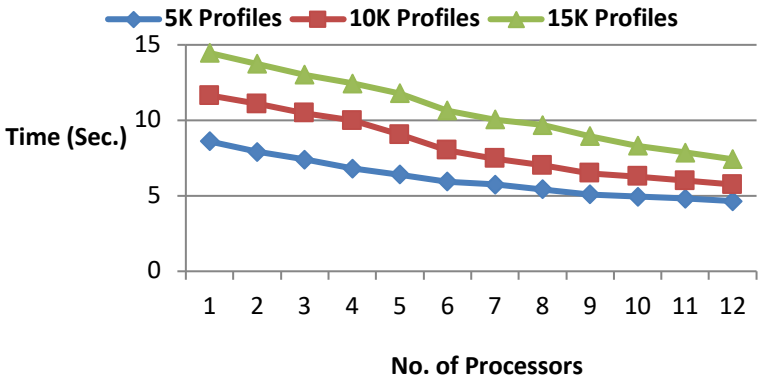




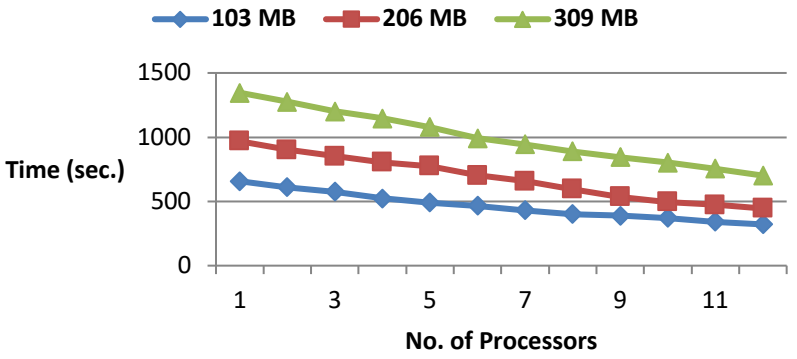


Case Study I: Scalability Analysis

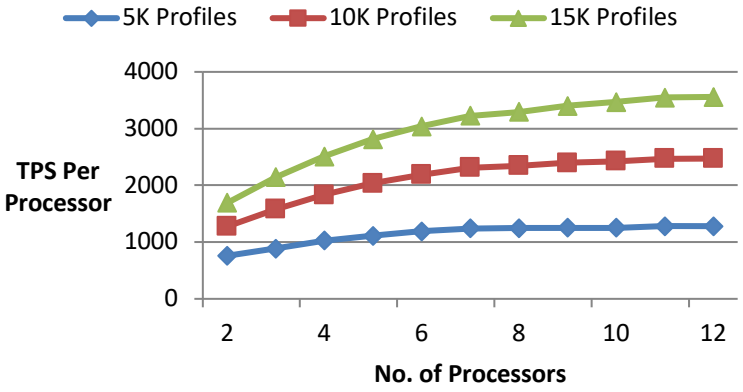
CFDRA Scalability Analysis by varying the no. of profiles and no. of processors



EUR Scalability Analysis by varying the no. of profiles and no. of processors

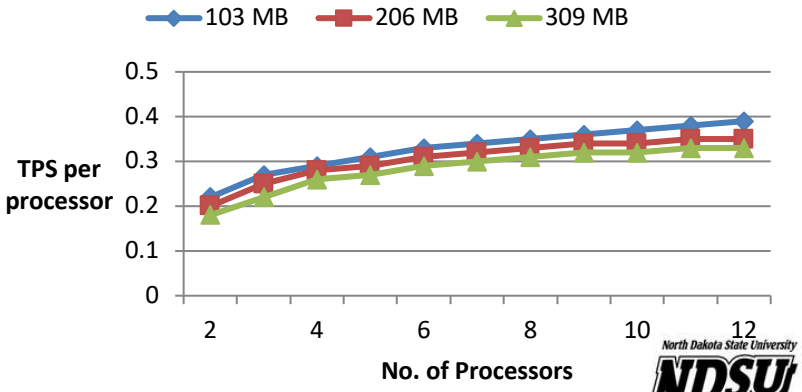


CFDRA Transactions per second per processor



TPS=No. of profiles compared

EUR Transactions per second per processor



TPS= Amount of data in MB

Case Study II: Health Insurance Plan Recommendation[1]

- Patient Protection and Affordable Care Act (PPACA)
  - Marketplaces
    - medical plans (78,000)[2]
    - dental plans (45,000) [3]
    - expected increase in near future
  - Private insurance providers
- Limited capabilities of the contemporary Web based tools
  - Challenges
    - Multi-faceted requirements
      - cost
      - coverage
    - Information filtering
      - difficult to find relevant information

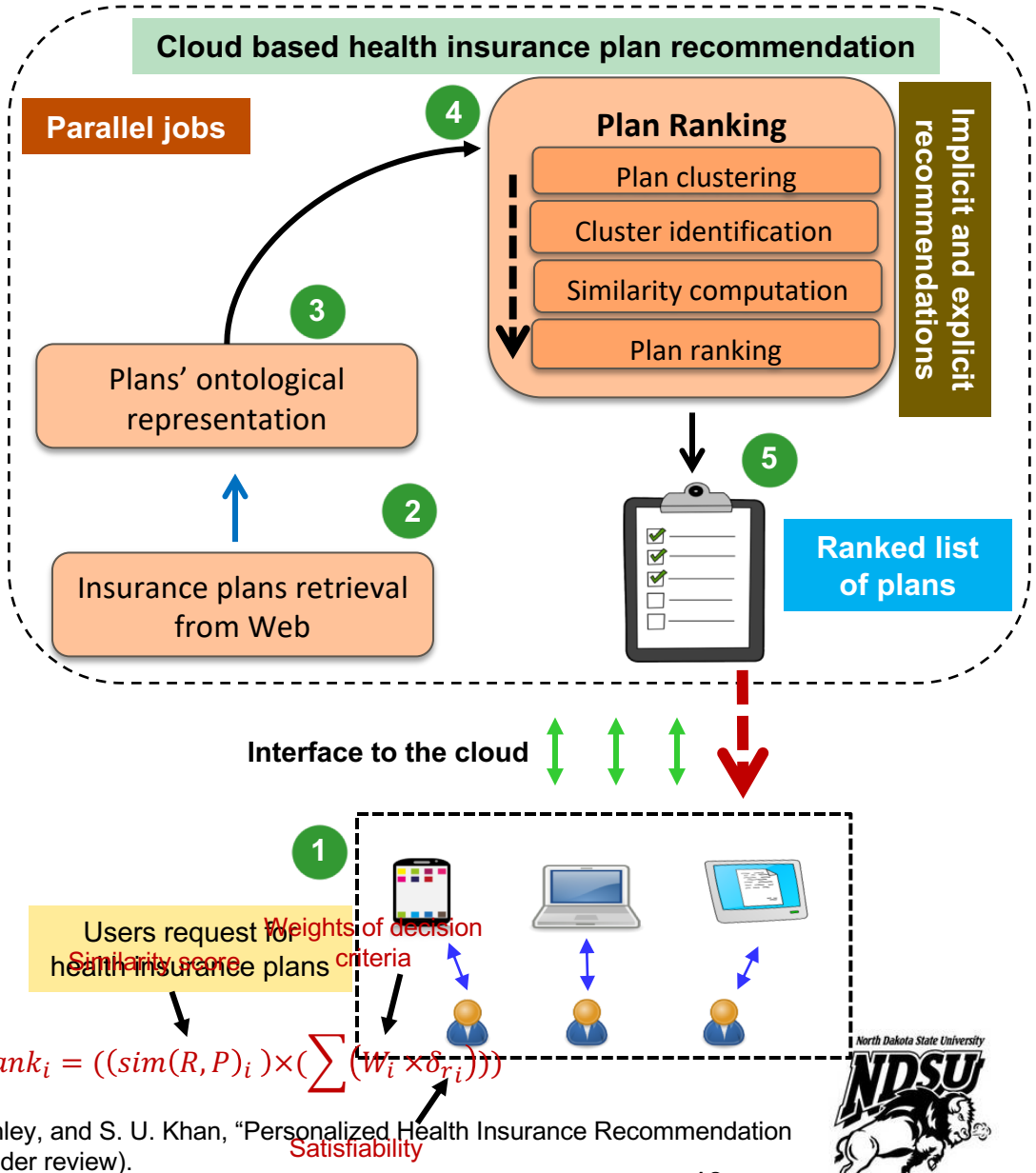
[1]A. Abbas, M. U. S. Khan, A. Yusoff, Y. Sadikaj, J. Ashley, and S. U. Khan, "Personalized Health Insurance Recommendation Services," *IEEE Transactions on Cloud Computing* (under review).
[2]QHP landscape individual market, https://data.healthcare.gov/dataset/QHPLandscape-Individual-Market-Medical/b8in-sz6k, 2015 (accessed on April 12, 2018).
[3]Dental plan information for individuals and families, https://www.healthcare. gov/dental-plan- information/, 2015 (accessed on April 12, 2018).

Case Study II: Health Insurance Plan Recommendation[1]

Key Accomplishments:

- Plans evaluation based on various criteria, such as premium, copay, deductibles, and out-of-pocket limit
- Implicit plan recommendations in the start (solution to cold start issue)
- Explicit plan recommendations based on user stated requirements
- Plans' clustering to minimize the number of comparisons
- A ranking methodology to rank the plans
- A methodology to avoid long-tail issue of recommender systems

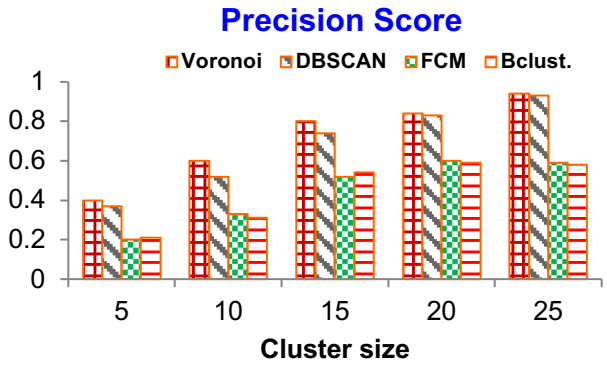
Implicit Recommendations	Explicit Recommendations
• Recommendations offered on first interaction with the system • Based on plan popularity • Initial popularity computation to overcome cold start	• Recommendations based on user stated requirements • Similarity between the plans and requirements • Ranking using Multi-attribute Utility Theory

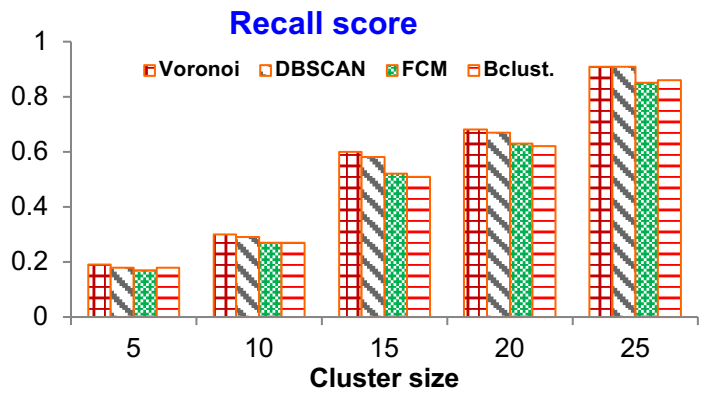


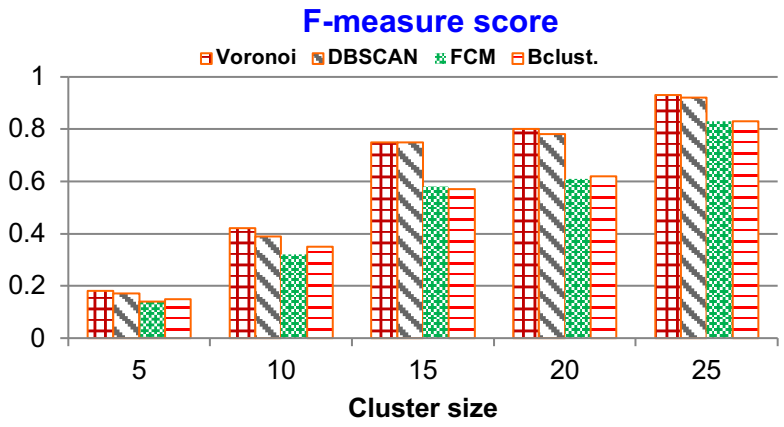
[1]A. Abbas, M. U. S. Khan, A. Yusoff, Y. Sadikaj, J. Ashley, and S. U. Khan, "Personalized Health Insurance Recommendation Services," *IEEE Transactions on Cloud Computing* (under review).

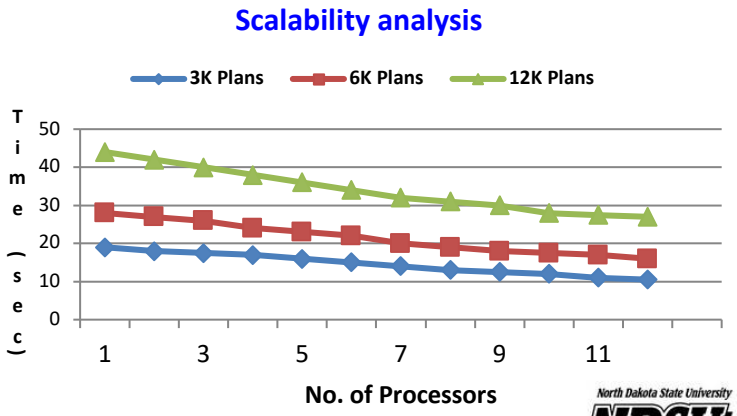
Case Study II: Experimental Results

DBSCAN— clustering based on density connected points
Fuzzy C Mean (FCM) — closeness to center
Voronoi—partitioning into cells based on ranking distance of plans
Bayesian Clustering (Bclust.)— cluster merging through statistical hypothesis test









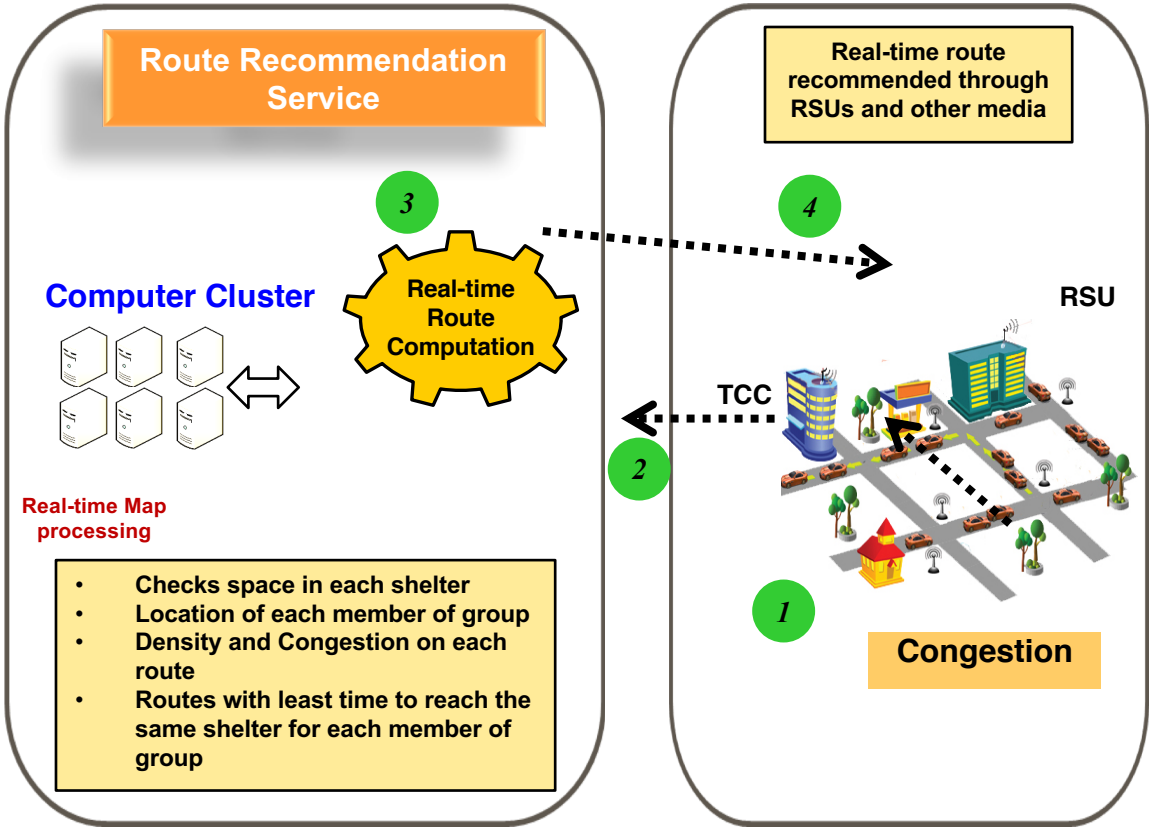
Case Study III: A Route Recommendation Service For Large-scale Evacuations[1]

Key Accomplishments:

- A scalable service capable of route recommendation during an emergency evacuation:
  - efficient traffic flows
  - leads to minimum congestion of the roads

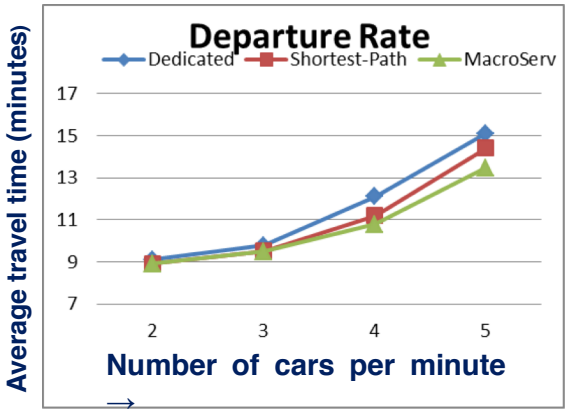
Challenges:

- Scalability:
  - big data graphs handling and partitioning
- Dynamic factors:
  - road congestions
  - road safety
  - shelter space

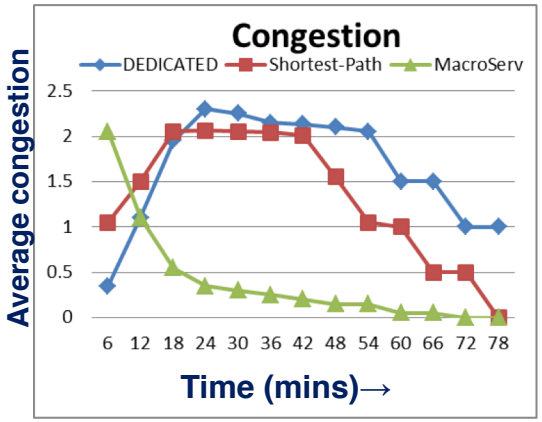


[1]M. U. S. Khan, O. Khalid, Y. Huang, F. Zhang, R. Ranjan, S. U. Khan, J. Cao, K. Li, B. Veeravalli, and A. Zomaya, "MacroServ: A Route Recommendation Service for Large-Scale Evacuations," *IEEE Transactions on Services Computing*, vol. 10, no. 4, pp. 589-602, 2017.

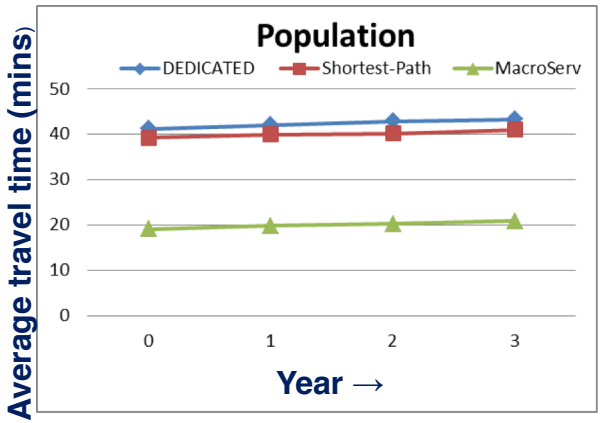
Case Study III: Experimental Results



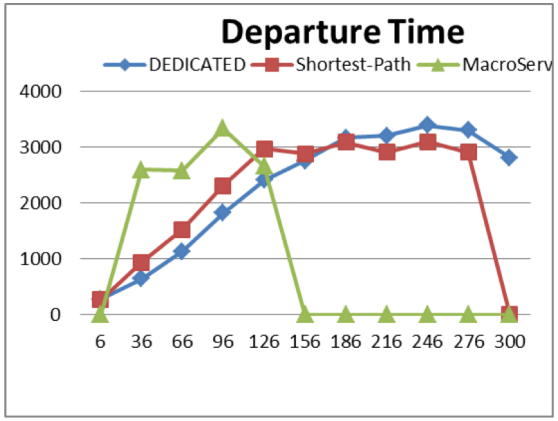
Average travel times with varying number of departing vehicles from each intersection



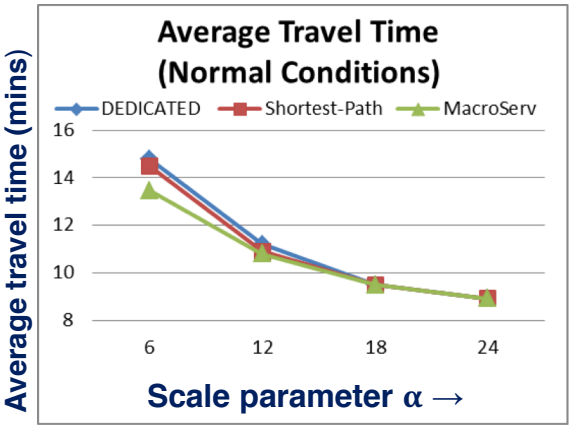
Average congestion with respect to time with damaged road network

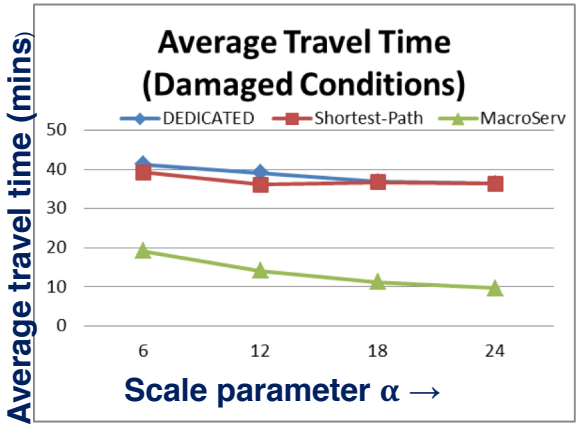


Effect of population increase in future 3 years on average car travel time on damaged network



Average evacuations per minute

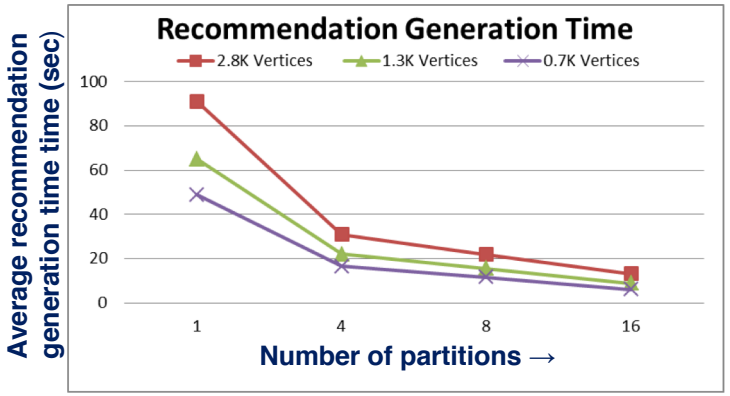


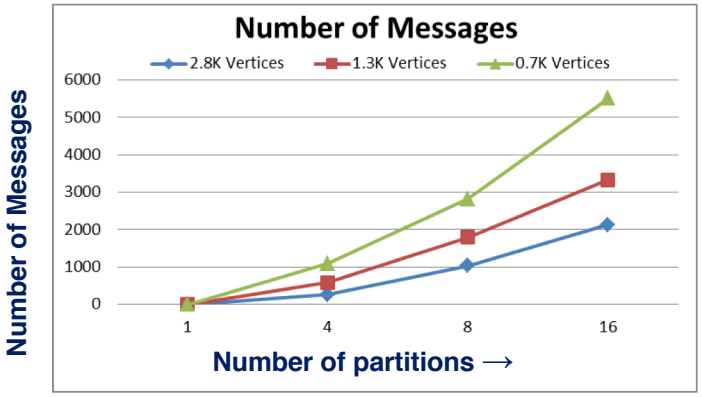


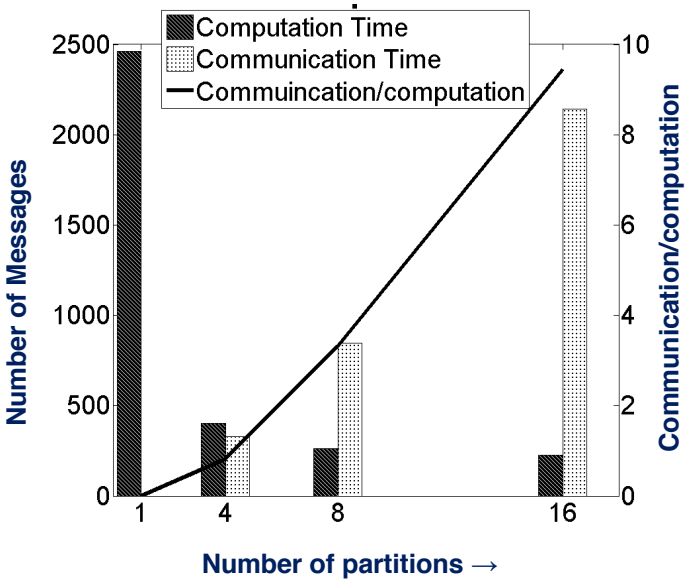
Effect of road damage by varying departure time (scale parameter α of Weibull Distribution)

Case Study III: Experimental Results

- Doubling the size of the region increases the recommendation generation time by an average of 26%.
- The increase in single processor results in decrease in the recommendation generation time by an average of 9%
- Doubling the size of the map decreases the average number of vehicle crossing from one zone to another by 76%.







Future Work

- Case Study I:
  - Identification of health experts from the same geographical area where enquiring users reside
  - Identification of fake twitter profiles through tweet analysis
- Case Study II:
  - Insurance plan recommendation through existing users characteristics
- Case Study III:
  - Considering additional parameters for emergency evacuations:
    - Drivers' behavior
    - Evacuees' compliance to the recommended routes