Distant Authorities

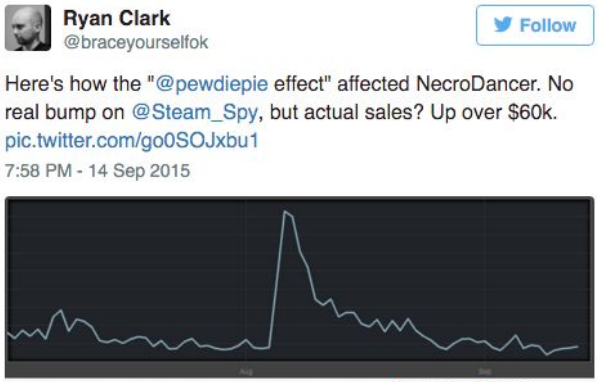
Anas Al-Khatib
Jun-Duo Chen
Aymeric Grail

Research Topic

Motivation:
Information drift over distance

Subject:
Distance from topical authority vs probability of acting on recommendation.

Inspired by discussions such as:
“the pewdiepie effect”

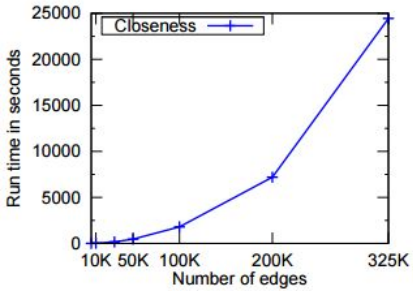


Papers: 1. Centralities in Large Networks: Algorithms and Observations [1]

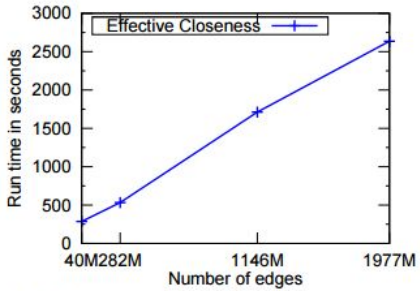
Centrality measures

- Effective closeness
- LineRank

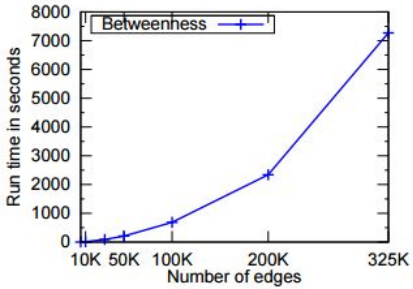
Parallel algorithms



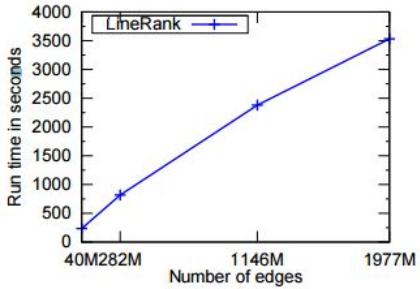
(a) Closeness: time vs. edges



(b) Effective Closeness: time vs. edges



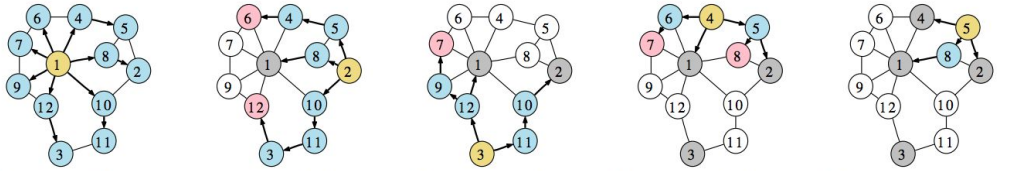
(d) Betweenness: time vs. edges



(e) LINERANK: time vs. edges

Papers: 2. Fast Exact Shortest Path Distance Queries on Large Networks [...] [2]

- Answer “Distance Queries” on large graphs with hundreds of millions of edges in microseconds
- Precompute distance labels of vertices by performing BFS from every vertex.
- Main addition pruning:



(a) First BFS from vertex 1. We visited all the vertices.

(b) Second BFS from vertex 2. We did not add labels to five vertices.

(c) Third BFS from vertex 3. We only visited the lower half of the vertices.

(d) Fourth BFS from vertex 4. This time we only visited the higher half.

(e) Fifth BFS from vertex 5. The search space was even smaller.

Figure 1: Examples of pruned BFSs. Yellow vertices denote the roots, blue vertices denote those which we visited and labeled, red vertices denote those which we visited but pruned, and gray vertices denote those which are already used as roots.

  - Reduces search space and size of labels
  - Can perform 32 or 64 BFS simultaneously by using bitwise operations
  - Based on the notion of distance labeling or distance-aware 2-hop cover.

Papers: 3. Finding Local Experts from Yelp Dataset [3]

Goal: Finding Local Experts from Yelp

1. Use of a random forest classifier to identify topical authorities
2. Combination of a map-reduce algorithm and a Gaussian Mixture Model to determine the locations a user has been living in
3. Determining a distance threshold under which a topical authority is considered a local authority

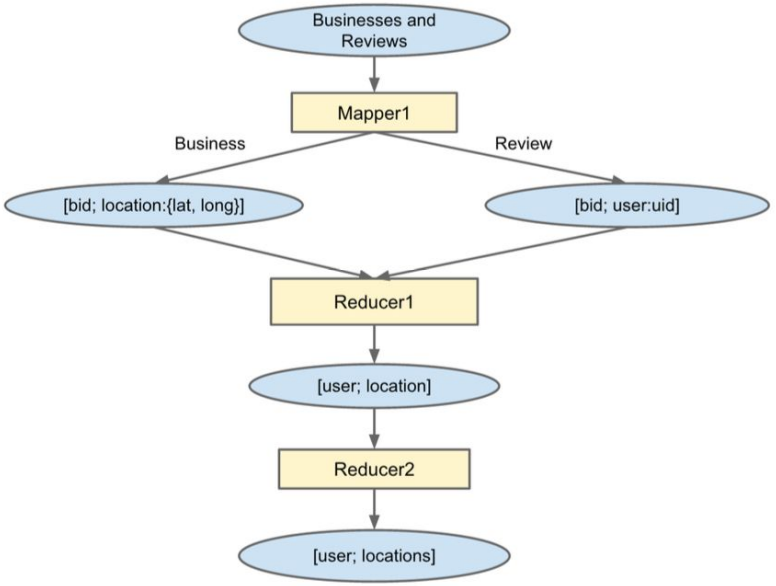


Figure 3.1: Map Reduce for user locations

Data

2.2M
Reviews

552K
Users

77K
Businesses

3.5M
Edges

https://www.yelp.ca/dataset_challenge

Research Goal - Assumption

Starting Assumption (what):

The smaller the social distance between a user and a topical authority, the more likely this user is to act upon the authority's recommendation.

Research Goal - Methodology

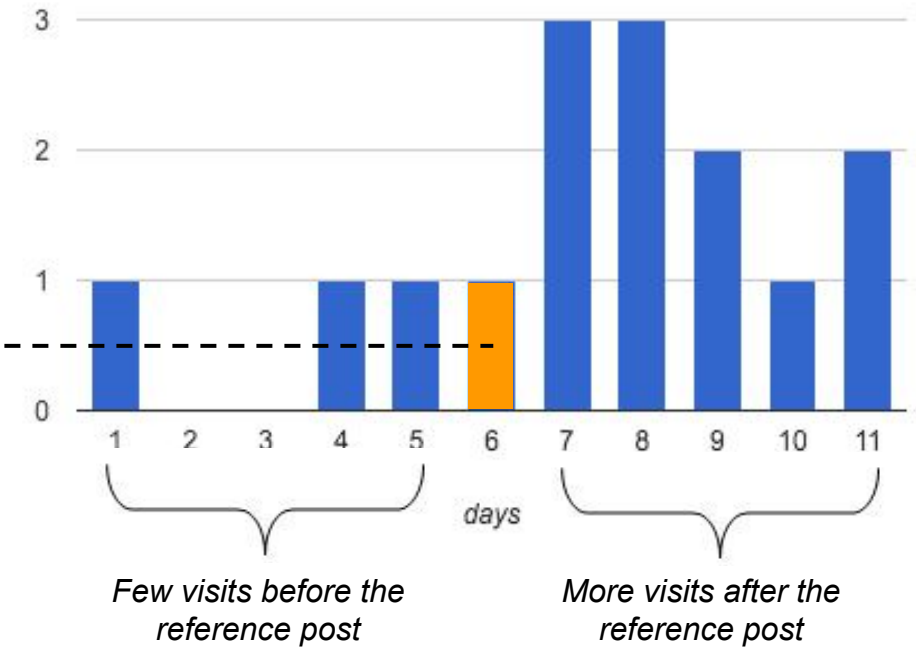
Methodology (how):

Compare the number of reviews before and after an authority's reference post. Then look at the distribution of those reviews depending on the reviewer's social distance to the authority.

Expected Impact of a Topical Authority

Scenario:
A topical authority gives 5 stars to a business.
We expect users at a close distance to be positively impacted

Topical Authority writes a 5 stars Reference Post



The Algorithm

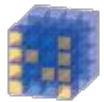
1. Determining Topical Authorities:
   - Classifier
   - Elite users

2. Calculating the distance between users and topical authorities

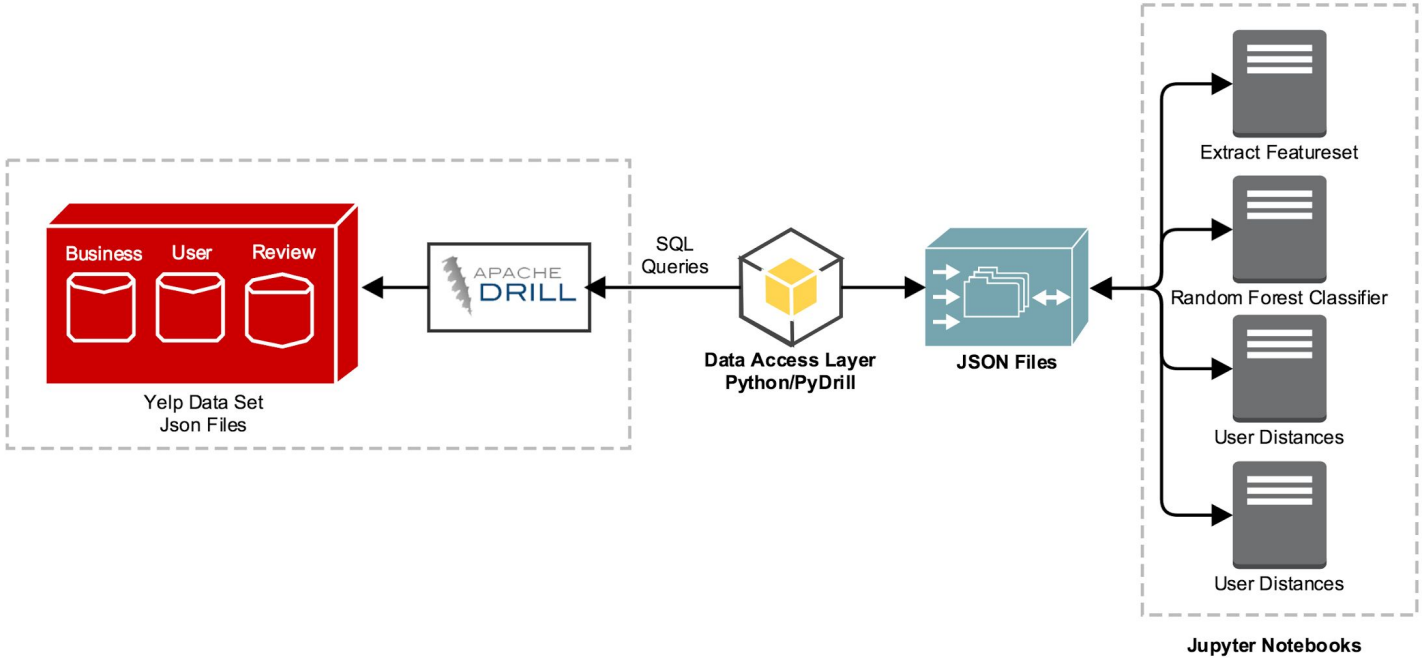
3. Determining the reviews distribution

Tools

- Apache Drill: Data Mining
- Python: Scripting language
- Jupyter Notebook: Python notebooks
  - Numpy
  - Pandas
- Pydrill: Python drivers for Apache Drill
- SciKit-Learn: Classifier



Workflow

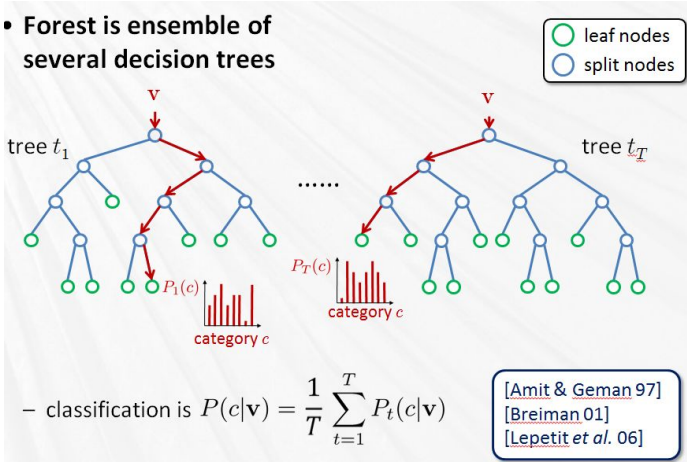

Business
User
Review
Yelp Data Set
Json Files
APACHE
DRILL
SQL
Queries
Data Access Layer
Python/PyDrill
JSON Files
Extract Featureset
Random Forest Classifier
User Distances
User Distances
Jupyter Notebooks

Determining Topical Authorities
Random Forest Classifier [3]

Train a classifier to identify experts vs. non-experts

Decision trees are commonly used for machine learning tasks, however they are seldom accurate and tend to overfit to the data.

“**Random Forests** are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance”



Determining Topical Authorities
Random Forest Classifier [3]

Extract features from dataset, used as input to the Random Forest Classifier

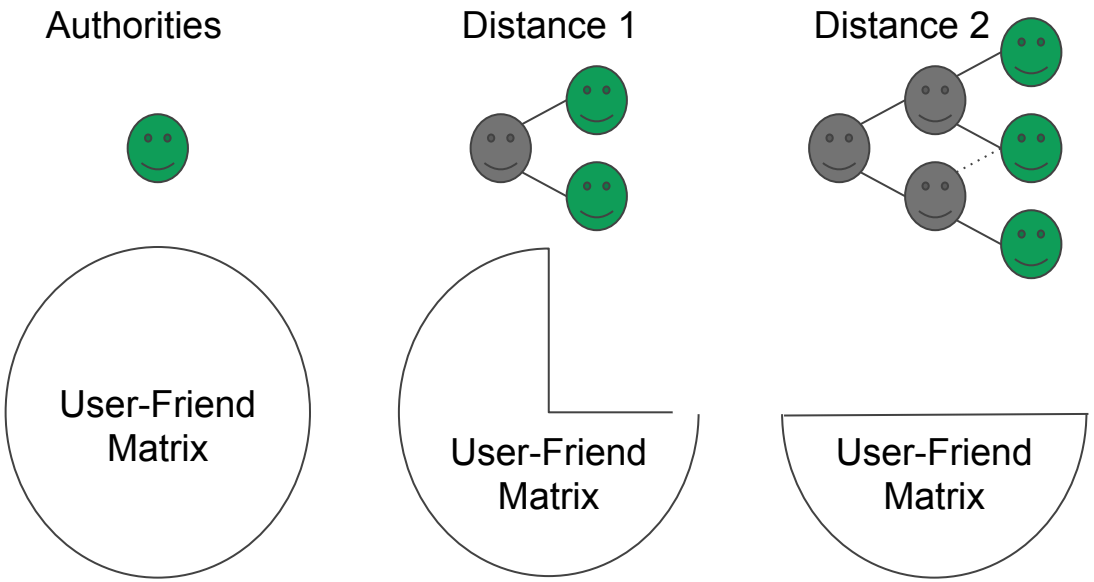
	categories_avg_rating	categories_biz_count	categories_reviews	cool	elite_user	funny	months_yelping	std_dev_rating	total_reviews	useful	user_id
0	5.000000	1	1	0	not_elite	0	17	0.000000	4	0	-BjekrEUhweLVyrjVStMgg
1	5.000000	1	1	0	not_elite	0	13	0.000000	6	1	XQiaNhoZbdxaN3pJIBADvw
10	5.000000	1	1	0	not_elite	0	16	0.000000	2	0	RCpt3acT2ZU4THf8pXdltg
100	5.000000	1	1	0	not_elite	0	38	0.000000	3	0	GjsRlhydavgl47QN1aiaPg
1000	4.000000	1	1	1	not_elite	1	62	0.000000	23	2	y-f7SiEV-2O1Q0vtz8-i1g
10000	1.500000	4	4	1	not_elite	4	54	0.866025	37	13	w98TrmskPMlMQLpz-Z8EVg

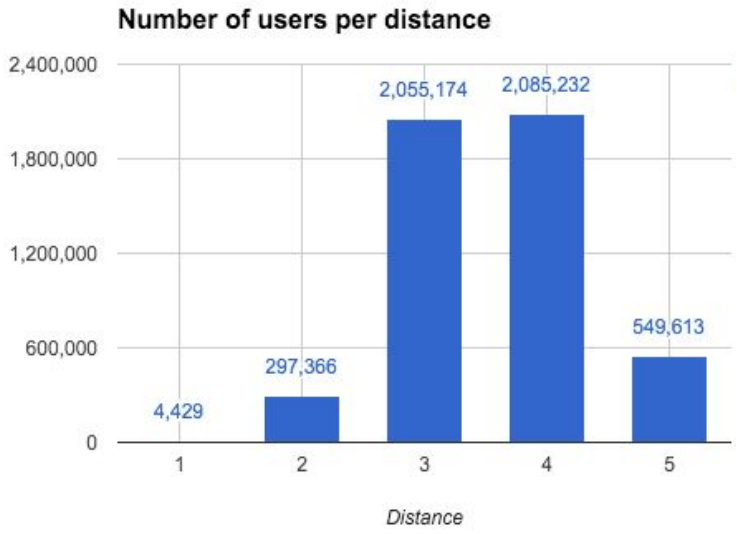
Train 67%	Test 33%

- Cross Validation Grid Search [5]
- Random Forest Classifier

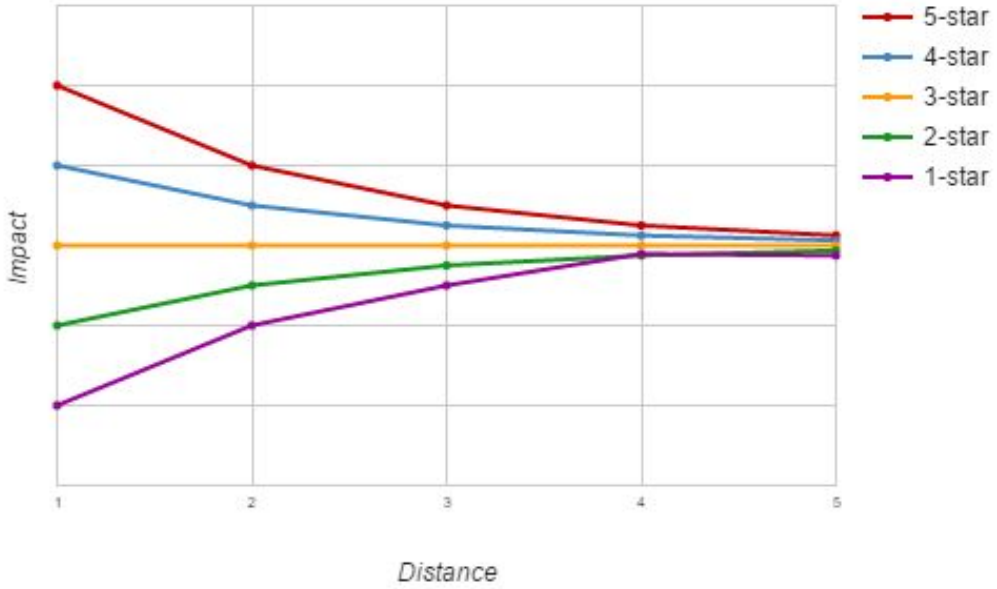
	precision	recall	f1-score	support
elite	0.71	0.32	0.44	2098
not_elite	0.90	0.98	0.94	13767
avg / total	0.88	0.89	0.87	15865

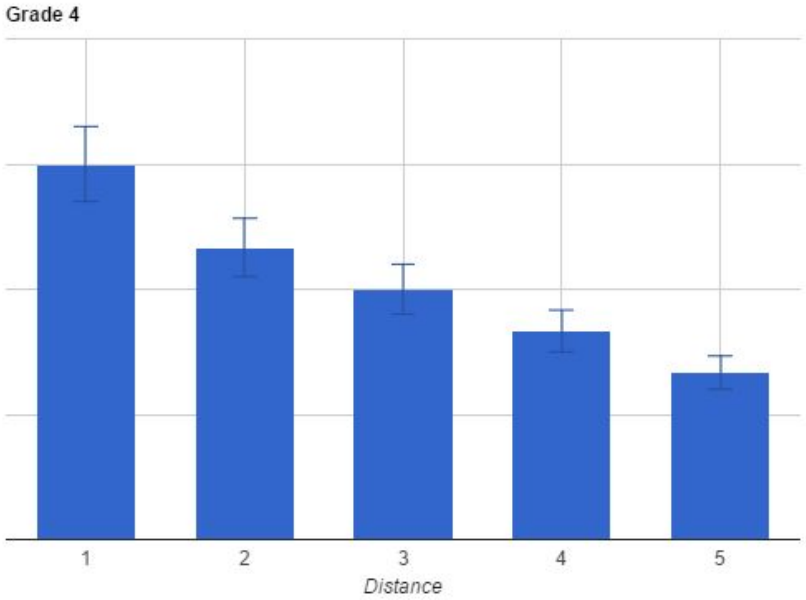
Calculating the Distance



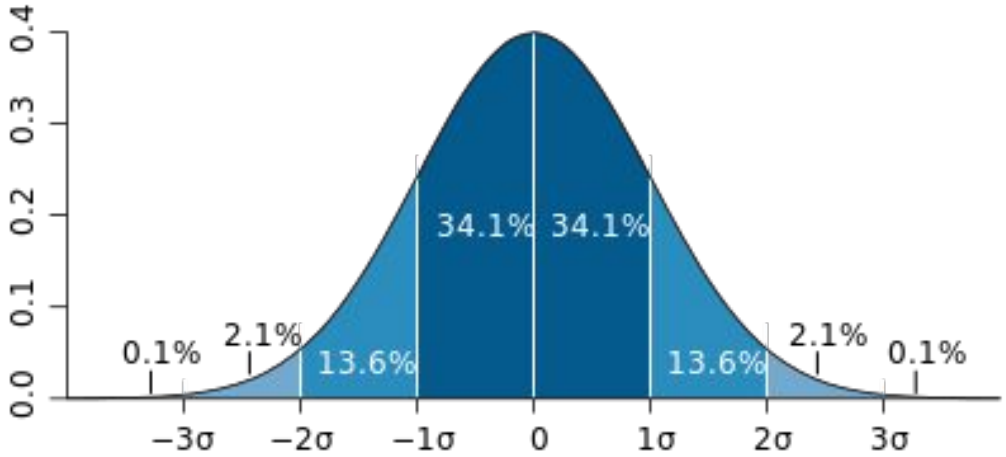


Results - Hypothesis



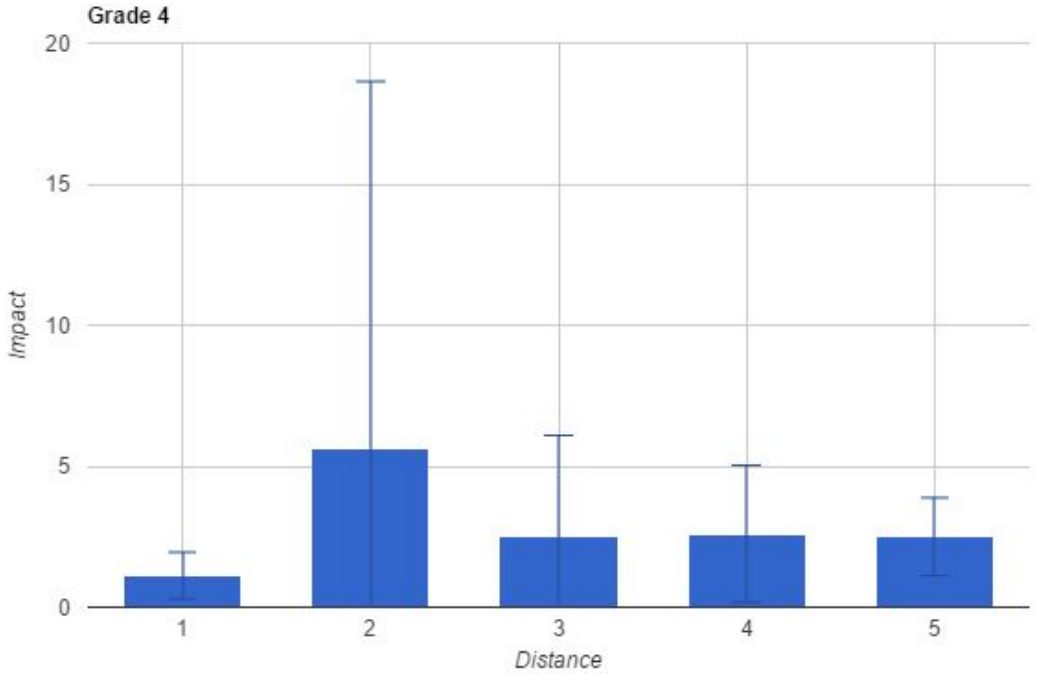


Results - Standard deviation



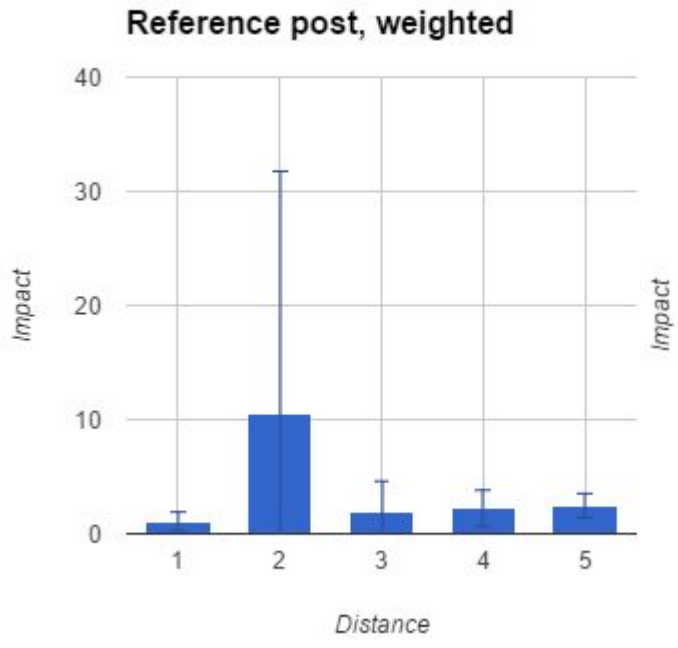
Source: https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Standard_deviation_diagram.svg
By Mwtoews - Own work, based (in concept) on figure by Jeremy Kemp, on 2005-02-09, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=1903871

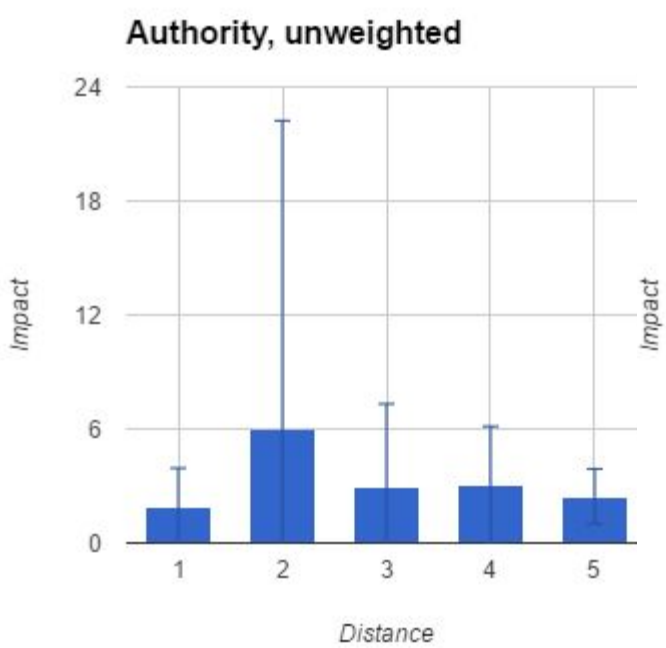
The Results - Initial

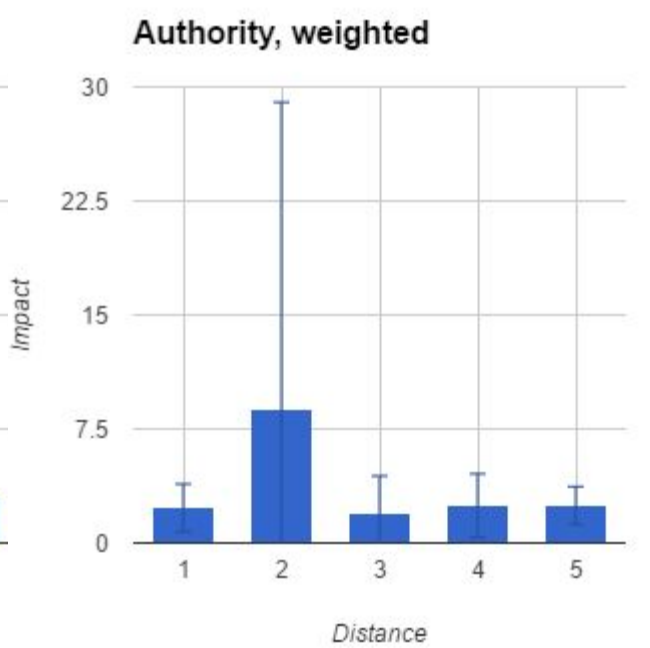


Grade	Distance	Impact	Sigma
1	2	2.914453	2.311739
1	3	3.699966	4.786833
1	4	1.461111	0.81174
1	5	1.4375	0.657489
2	2	1.826141	2.644155
2	3	1.266526	0.66747
2	4	1.889069	1.754287
3	1	2.6	2.046678
3	2	2.698464	3.568751
3	3	6.381141	13.408655
3	4	8.122917	15.737784
...	...	...	...
5	2	1.673929	2.684528
5	3	2.619243	3.011214
5	4	4.621212	9.788128
5	5	2.2	0.447214

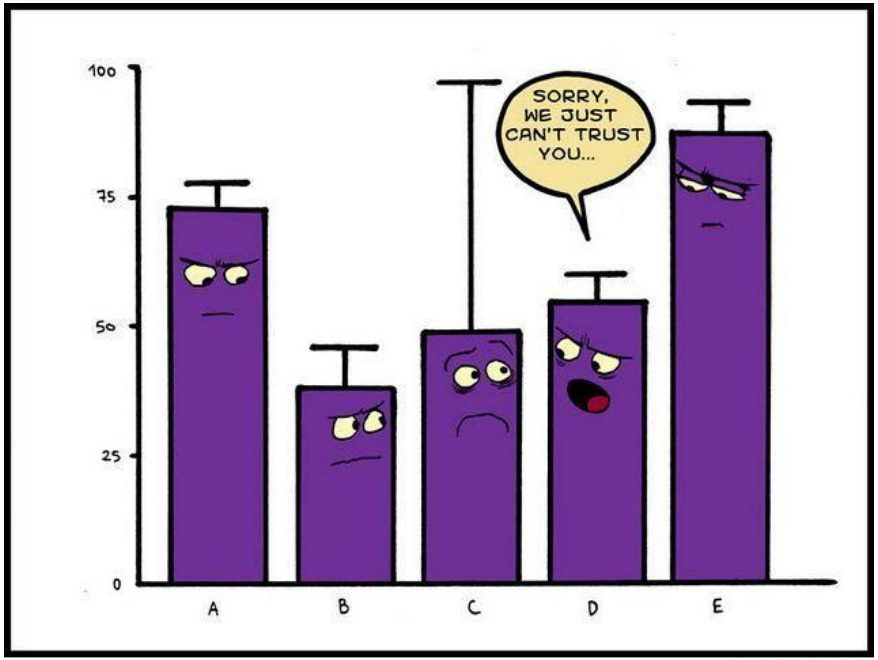
The Results - Alternate Processing







Conclusions



Limitations

- Maybe overall grade is more representative of an authority’s influence than the number of reviews.
- Visits data not provided
- The model assumed a normal distribution (relevance of standard deviation)
- Impact of multiple authorities

Future Work

- Topical authority selection algorithm
- Impact as ratio
- Consider Tips and physical visits to restaurants
- Effect of time on impact

References

1. Kang, U., Spiros Papadimitriou, Jimeng Sun, and Hanghang Tong.
"Centralities in Large Networks: Algorithms and Observations."
Proceedings of the 2011 SIAM International Conference on Data Mining(2011): 119-30. Web.
(link: http://www.cs.cmu.edu/~ukang/papers/CentralitySDM2011.pdf)

2. Akiba, Takuya, Yoichi Iwata, and Yuichi Yoshida.
"Fast Exact Shortest-path Distance Queries on Large Networks by Pruned Landmark Labeling."
Proceedings of the 2013 International Conference on Management of Data - SIGMOD '13 (2013): n. pag. Web.

3. Jindal, Tanvi. "Finding Local Experts from Yelp Dataset." Diss. U of Illinois at Urbana-Champaign, 2015.
IDEALS @ Illinois. 27 Apr. 2015. Web. 07 Feb. 2016.
(link: https://www.ideals.illinois.edu/handle/2142/78499)

4. Yelp Data Set Challenge: https://www.yelp.com/dataset_challenge

5. Lam, Diana. "Predicting Yelp's Elite." Predicting Yelp's Elite. N.p., 28 Feb. 2016. Web. 02 Apr. 2016. <http://dianalam.github.io/2016/02/28/yelp-classification.html>.