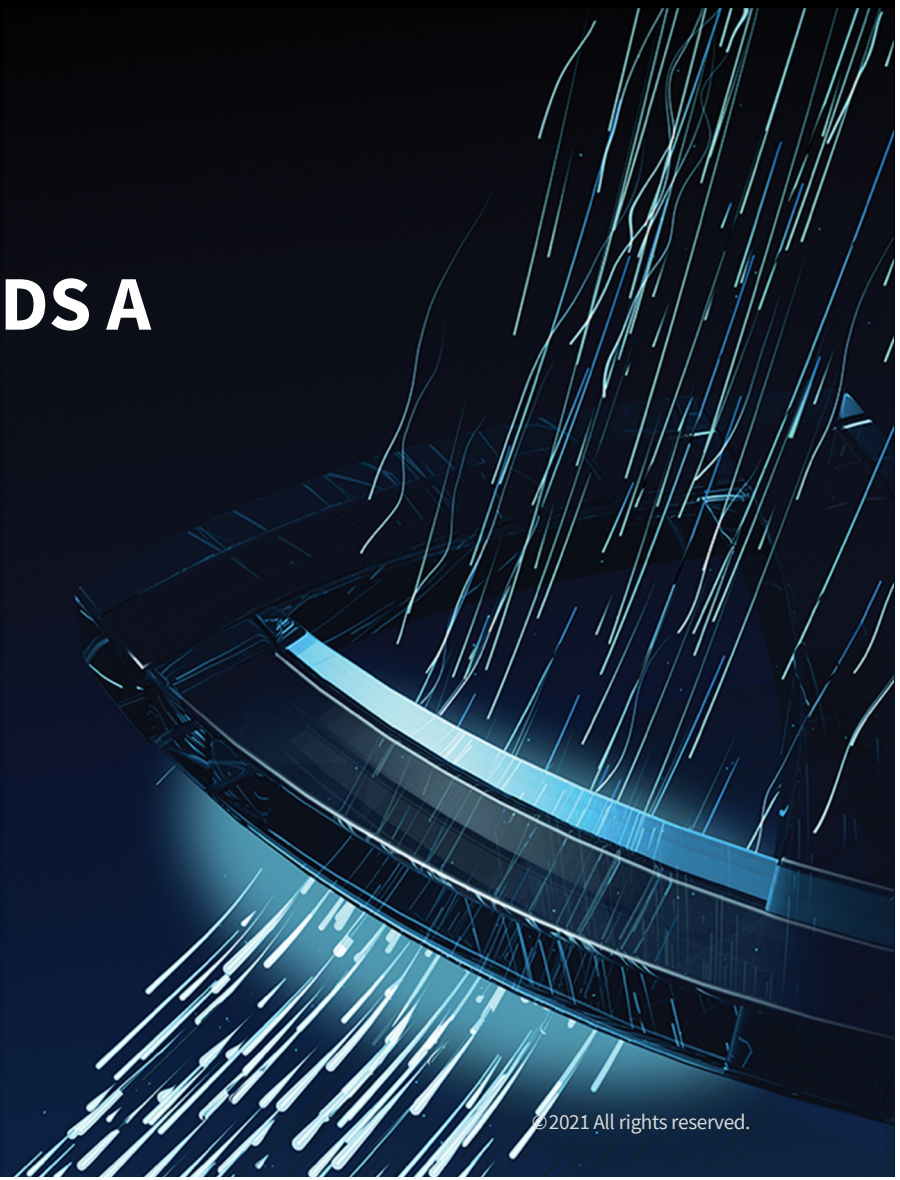
MULTIGRAPH APPROACH TOWARDS A SCALABLE, ROBUST LOOK-ALIKE AUDIENCE EXTENSION SYSTEM

Ernest Kirubakaran Selvaraj
Tushar Agarwal
Nilamadhaba Mohapatra
Swapnasarit Sahu



ZEOTAP

Gartner COOL VENDOR 2020

2019 WINNER

CDP INSTITUTE

Leader SPRING 2021

AGENDA

- Background
- Proposed Method
- Experiments
- Recap & Future Work

ZEOTAP

Background

ZEOTAP

EXPAND ADVERTISING AUDIENCE, TARGET PEOPLE WHO ARE MOST SIMILAR TO CONVERTED USERS

WHAT
IS A LOOK-ALIKE MODEL?

A model that provides an extended set of target customers who are most like the seed set

Seed set: Set of customers who have responded positively to the ad campaign

HOW
IS IT BUILT?

By finding users in the global user set who share similar demographic and behavioral profiles with the seed set

Hypothesis: Users with similar profiles act-alike

WHY
IS IT CHALLENGING?

Advertising data is getting increasingly complex, high-dimensional and extremely sparse

Nature of business demands low latency

Proposed Model

- Data
- Multigraph
- Scoring Methodology

ZEOTAP

DATA

- User profile data available for look-alike modeling can be broadly grouped into Demographics and Behavioral data

DEMOGRAPHICS

Users' age, gender, location, income group, etc.

Many of demographic features are high-cardinal in nature e.g., user's location/city

Data is usually sparse – Only a handful of information is known about every user with certainty

Behavioral Data

Product purchase data, mobile app usage data, purchase intent & interest data etc.

Single user can have multiple values for a feature

Cardinality is usually very high

MULTIGRAPH

Model Objective:

- Accuracy: Given a seed set of users, find users with most similar profiles
- Latency: Extended user set must be retrieved fairly quickly

Solution: Offline graph building

- Build a k-nearest neighbor (KNN Graph) user-user multigraph where every user is connected to k most similar neighbors
- Separate graphs are built for demographic features and every behavioral features such as mobile app usage, interest & intent, etc.,
- The graphs are merged to form a multigraph, each graph having its own edge type
- Each graph is deterministic and provides accurate neighbors
- Since the graph building is offline, extension is fast as it involves only retrieval of neighbors from the multigraph

MULTIGRAPH

Building deterministic graphs

- Building KNN Graphs has a complexity of $O(n^2)$ and hence not feasible with datasets with millions of users
- We use an iterative method NN-Descent [1] which has an empirical cost of $O(n^{1.14})$ to build the KNN Graphs
- NN-Descent algorithm is guaranteed to converge when the distance measure used is a complete metric and a large enough ‘k’ is used
- We use different approaches and metrics to build the graph for demographic variables that are categorical in nature and other behavioral variables that can take multiple values per user

MULTIGRAPH

Demographic Graph

- Demographic features are categorical and high cardinal in nature and an appropriate distance metric must be defined to build the KNN Graph
- If we consider each feature such as gender, location, etc., as a dimension, Euclidean distance between any two users can be calculated if we can define a measure of how far any two categories are apart in each feature dimension
- We use Hellinger Distance as a proxy for the relative distance between any two feature values
- Hellinger Distance between two feature values j and k belonging to a feature f is given by

$$HD(f_j, f_k) = \sqrt{1 - \sum_{\substack{l \in f' \\ \forall f' \in F, f' \neq f}} \sqrt{p_{sj}(f' = f'_l) p_{sk}(f' = f'_l)}}$$

- Once we have the Hellinger Distance between the feature values, we can calculate the Euclidean distance between any two users and employ it as a distance metric in the NN-Descent algorithm

MULTIGRAPH

Behavioral Graphs

- We build a separate graph for every behavioral feature such as mobile app usage data, interest & intent data, etc.,
- One characteristic feature of behavioral data is that users can have multiple values for any feature. For e.g., if we take mobile app usage data, every user may have anywhere between 5 to 200 apps on their mobiles
- These features are very high dimensional, and the graph won't converge to the global minimum if we keep the feature dimension large
- We first learn an embedding for every feature value in a behavioral feature by factorizing the global co-occurrence matrix
- If the embeddings are n-dimensional and every user can have a maximum of m feature values ($m<n$), then each user can be represented as a mxn matrix. In other words, every user occupy a m-dimensional sub-space inside the n-dimensional embedding space

MULTIGRAPH

Behavioral Graphs

- To calculate the distance between two user subspaces, we use a modified Chordal distance

$$d(U,V) = \sqrt{max(m,n) - \sum_{i=1}^{m}\sum_{j=1}^{n}(u_i^T v_j)^2}$$

- Here U and V are the matrix representation of two users and u and v are the orthonormal bases of U and V
- The behavioral graph is built using this modified Chordal distance as a metric

MULTIGRAPH

Scoring

- When a seed set is received from a campaign, set of potential look-alike users is formed by selecting the neighbors of the seed set users in the global multigraph
- Then the candidate look-alike users need to be scored based on their likelihood that they belong to the seed set
- The relative importance of a candidate user depends on his features and the structure of the graph
- Importance due to features is calculated using the weighted Information Value between the seed set distribution and the global user distribution

$$Weighted\ IV(c \in C) = \sum_{f \in F} \sum_{j \in f} p_s(k)\ x_{c,f=j}\ IV(f)$$

- For every candidate user, the number of seed users in the neighborhood and the number of edges with seed user is also an indicator of importance

MULTIGRAPH

Scoring

- The final score of a seed set user is calculated as

$$Score(c \in C) = N_c \, E_c \sum_{f \in F} \sum_{j \in f} p_s(k) \, x_{c,f=k} \, IV(f)$$

- Here, N_c is the number of seed set users in the neighborhood of candidate user c and E_c is the number of edges between the candidate user and the seed set user
- Once the candidate users are scored, top candidate users are given as extended audience

Experiments

- Online A/B Tests
- Seed Set Recovery

ZEOTAP

ONLINE A/B TESTING

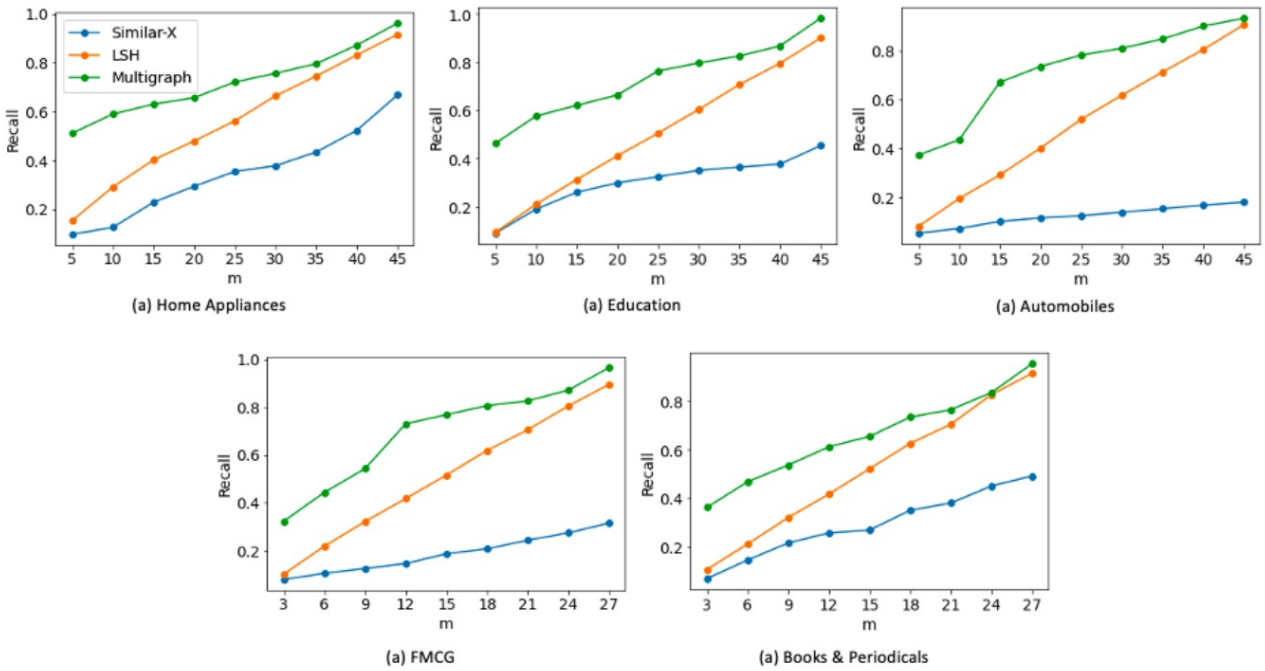
- We ran a set of online campaigns to compare the performance of our model and the LSH [2] and Similar-X models [3].
- Table 1 shows the Click-Through Rates achieved using the three models on various campaigns
- Our approach gave better performance in all the 5 campaigns

Campaign	Baseline	Similar-X	LSH	Multigraph
Appliances	1.85%	4.05%	7.73%	**8.02%**
Education	1.25%	3.21%	7.07%	**7.38%**
Automobiles	1.36%	2.67%	7.77%	**8.67%**
FMCG	1.68%	2.20%	3.10%	**3.25%**
Books	2.21%	2.45%	4.15%	**5.38%**

Table 1: CTR Rates for various campaigns

ONLINE A/B TESTING

- We also evaluated how far each model can recover the original seed set by extending on a subset of the seed set from the campaigns
- In this seed set recovery task, our approach gave better results especially when the extension is small

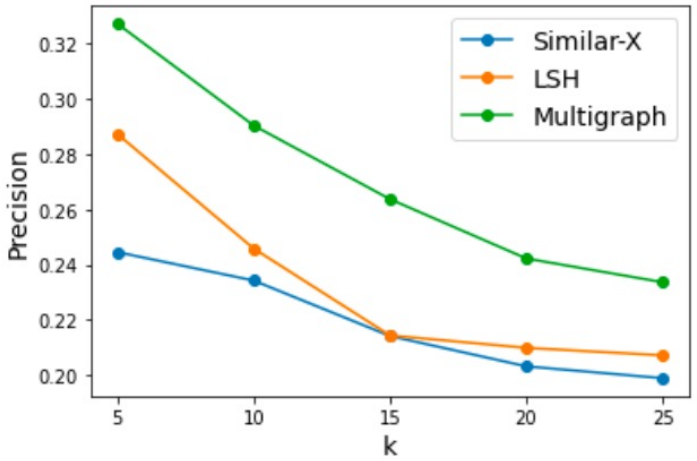


Recall at different rates of m for the 5 campaigns

ZEOTAP

SEED SET RECOVERY

- In the second set of experiments on Adform Click Prediction Dataset [4], we ran a seed set recovery experiment.
- The data contains clicks recorded and a set of anonymized variables
- We take a random subset of clicks from the dataset consisting of 50,000 records and extended the records to assess the precision of extension using the three models
- Here again, multigraph model gave better performance than the other two models



Precision@k for different k values on Adform Data

Conclusion

ZEOTAP

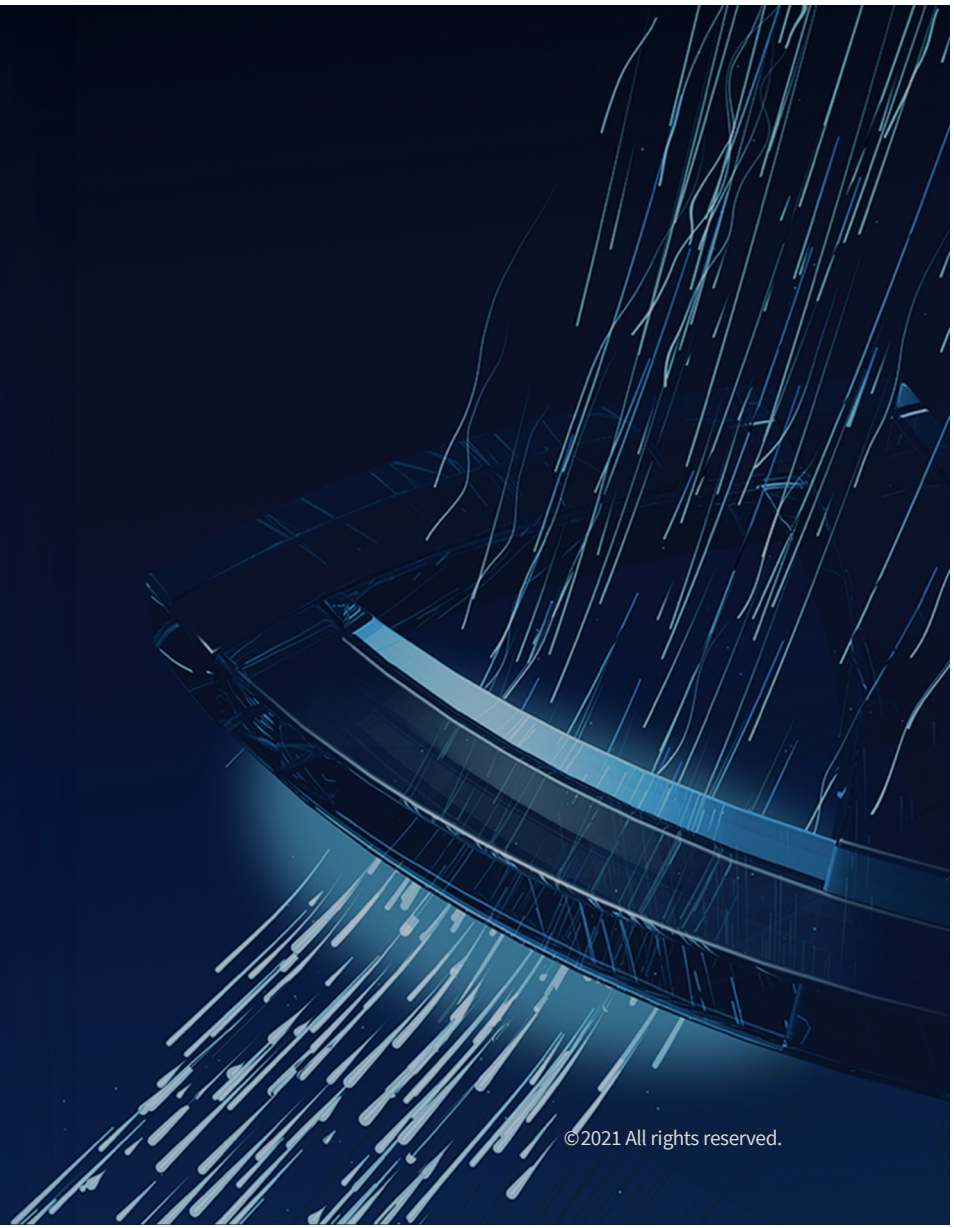
CONCLUSION

- We propose a multigraph look-alike audience extension system
- Allowing different graphs for different categories makes neighborhood searches more robust and accurate
- The model is computationally efficient during prediction time as it involves only retrieval and scoring of candidate users
- The model can be easily scaled to millions of users
- Real-world experiments show that our model achieves better CTR rates than existing audience extension models
- Our model performs well in segment recovery tasks as well due to the deterministic graphs

Future Work

- Account for changes in short-lived user data such as purchase intent information

THANK YOU



ZEOTAP

REFERENCES

- Wei Dong, Charikar Moses, and Kai Li. 2011. Efficient k-nearest neighbor graph construction for generic similarity measures. In Proceedings of the 20th interna- tional conference on World wide web. 577–586. [1]
- Qiang Ma, Musen Wen, Zhen Xia, and Datong Chen. 2016. A sub-linear massive-scale look-alike audience extension system. In Proceedings of the 5th International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications. [2]
- HaishanLiu,DavidPardoe,KunLiu,ManojThakur,FrankCao,andChongzheLi. 2016. Audience expansion for online social network advertising. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 165–174. [3]
- Enno Shioji. 2017. Adform click prediction dataset. https://doi.org/10.7910/DVN/TADBY7 [4]