PlayStore Apps Analysis

Gopal Ramesh Dahale (11840520)
Network Science (CS552)

Department of EECS
Indian Institute of Technology Bhilai

May 2, 2021

Outline

1. Introduction
2. Related Works
3. Dataset
4. Elementary Analysis of Original Network
5. Elementary Analysis of Sampled Network
6. Community Detection
7. Survey
8. Conclusion

Introduction

- Focus on analyzing the apps from the Google Play store that provides a wide range of data features (price, rating, etc.)
- We choose Google Play Store over all other markets because of its growing success and recent rapid growth. 96.7 percent of apps are free is one of the key reasons for the growth[1].
- We here explore the properties formed by the network of apps.

Figure: Google Play Store Logo[2]

[1] *Number of free apps on Google play store*. 2021. URL: https://www.statista.com/.

[2] https://zeenews.india.com/apps/mitron-app-suspended-from-google-play-store-2287704.html.

Outline

1. Introduction
2. Related Works
3. Dataset
4. Elementary Analysis of Original Network
5. Elementary Analysis of Sampled Network
6. Community Detection
7. Survey
8. Conclusion

- The majority of this approach was inspired by[3] and[4] description of how to construct a network from a list of Amazon
- We used their approach to create a network of Google Play Store apps.
- [5] helped us to decide how and which community detection algorithms to chose to analyse the network.

[3] Song and Zhao, "Survey of Graph Clustering Algorithms Using Amazon Reviews".

[4] Julian McAuley, Rahul Pandey, and Jure Leskovec. "Inferring Networks of Substitutable and Complementary Products". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. Sydney, NSW, Australia: Association for Computing Machinery, 2015, 785–794. ISBN: 9781450336642. DOI: 10.1145/2783258.2783381. URL: https://doi.org/10.1145/2783258.2783381.

[5] Shihui Song and Jason Zhao. "Survey of Graph Clustering Algorithms Using Amazon Reviews". In: (

Outline

1. Introduction
2. Related Works
3. Dataset
4. Elementary Analysis of Original Network
5. Elementary Analysis of Sampled Network
6. Community Detection
7. Survey
8. Conclusion

App Data

- Scraped from the official website of Google Play Store[6] using google-play-scraper[7]
- Over 61 categories. The category-wise distribution is shown in fig 2.

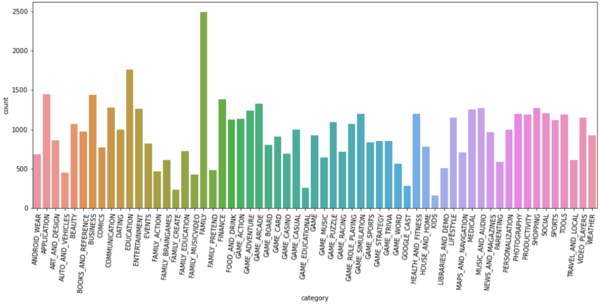

Figure: Category wise app distribution

- A total of 1,54,637 were scraped. 57,590 apps left after removing duplicates.

[6] Google. *Google Play Store*. https://play.google.com/store. 2008.

[7] Facundo Olano. *google-play-scraper*. https://github.com/facundoolano/google-play-scraper. 2019.

Reviews Data and Features

- Scraped using google-play-scraper[8].
- At most 1000 reviews per app.

App Data Features

- **AppId:** Unique Id for each app.
- **Price:** Price of an app. The currency is Indian Rupee.
- **scoreText:** Rating of app shown on playstore.
- **category:** Category of an app out of 61 possible categories.

Reviews Data Features

- Reviewer Id was not available, so we used **Reviewer's Name** as the edge parameter.
- For the pair of apps which a reviewer reviewed, we have the **scores** given by the reviewer as an attributes.

[8] JoMingyu. *Google-Play-Scraper*. https://github.com/JoMingyu/google-play-scraper.

Pre-Processing and Network Formation

- Chose Edge lists because of the simplicity they provide.
- Listed the apps reviewed by each reviewer and created a network.
- Node represents an app and link represents a pair of apps that a single reviewer has reviewed.

	Count	File Size
Nodes	57,588	22 MB
Edges	2,95,33,258	2.5 GB

Average degree: 967.0400

Graph Sampling

- Huge network. Time-consuming and computationally expensive.
- Pick a subset of vertices/ edges from the network.
- Random walk Induced Graph Sampling as implemented in[9] and studied in[10].

	Count	File Size
Nodes	11,453	4.3 MB
Edges	50,48,335	380 MB

Average degree: 881.5743

[9] Ashish Aggarwal. *Graph Sampling Package*. https://github.com/Ashish7129/Graph_Sampling.

[10] Jure Leskovec and Christos Faloutsos. "Sampling from Large Graphs". In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '06. Philadelphia, PA, USA: Association for Computing Machinery, 2006, 631–636. ISBN: 1595933395. DOI: 10.1145/1150402.1150479. URL: https://doi.org/10.1145/1150402.1150479.

Outline

1. Introduction
2. Related Works
3. Dataset
4. Elementary Analysis of Original Network
5. Elementary Analysis of Sampled Network
6. Community Detection
7. Survey
8. Conclusion

- Network density: 0.021.
- On a scale of 0 to 1, not a very dense network.
- This is the density of *whole* network, including disconnected components.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
Nodes	45798	2	2	2	2	2	2	3
Edges	22151492	1	1	1	1	1	1	3

- Density of the largest connected component is 0.021.
- Maximum degree: 4211 of a shopping app and then 3906 of a food/drink app.
- Degree exponent of $1.17 < 2$ i.e. anomalous regime.

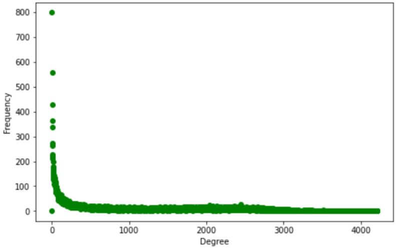

Figure: Degree Distribution of original network

Outline

1. Introduction
2. Related Works
3. Dataset
4. Elementary Analysis of Original Network
5. Elementary Analysis of Sampled Network
6. Community Detection
7. Survey
8. Conclusion

Degree Centrality

Connected component. Network density : 0.07. Not a very dense network.

App Id	Price	Rating	Category	Degree	Eigenvector
net.tsapps.appsales	Free	4.3	Shopping	1914	0.018
com.snapbreak.doors	Free	4.5	Game Puzzle	1885	0.017
com.technologies.subtlelabs.doodhvale	Free	4.4	Food and Drink	1883	0.02
com.randomvideochat.livevideochat	Free	3.9	Communication	1848	0.019
com.socialnetwork.metu	Free	4.9	Social	1827	0.019

Table: Top 5 apps with highest degree centrality

App Id	Price	Rating	Category	Deg	Eigen
passport.Size.Photo.Maker.Editor.Countries	Free	0	Productivity	3	0
com.kiroglue.lookalikemomordadsimilarityparents	Free	4.0	Parenting	3	0
com.photovideomaker.slideshow.videostatusmaker	Free	0	Family Music	5	0
com.medpresso.Lonestar.manotes	Free	4.6	Medical	5	0
com.tutioncentral.cmanoteslite	Free	4.6	Medical	5	0

Table: Top 5 apps with lowest degree centrality

It seems that the more the number of reviewers, the higher is the rating. Although, from table 2 the two medical happens to have significant rating even though they have less degree.

Degree Distribution

Fitting the distribution to power-law gave a degree exponent of $1.15 < 2$ i.e. again anomalous regime.

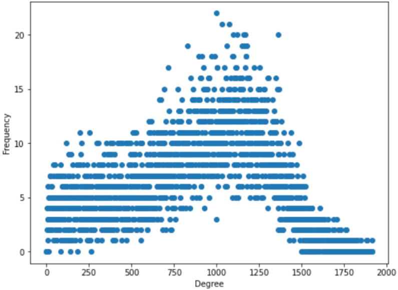

Figure: Degree Distribution of sampled network

Eigenvector Centrality

The maximum eigenvector centrality value is 0.02 for the network.

App Id	Price	Rating	Category	Degree	Eigenvector
com.technologies.subtlelabs.doodhvale	Free	4.4	Food and Drink	1883	0.02
com.randomvideochat.livevideochat	Free	3.9	Communication	1848	0.019
com.socialnetwork.metu	Free	4.9	Social	1827	0.019
com.u2opia.woo	Free	4.2	Application	1814	0.019
com.edudrive.exampur	Free	4.2	Education	1787	0.018

Table: Top 5 apps with highest eigenvector centrality

Very less on a scale from 0 to 1. Most of the apps with a high degree and eigenvector centrality are free apps and have a reasonable rating.

Outline

1. Introduction
2. Related Works
3. Dataset
4. Elementary Analysis of Original Network
5. Elementary Analysis of Sampled Network
6. Community Detection
7. Survey
8. Conclusion

Clauset-Newman-Moore greedy modularity maximization[12]

- Time complexity of $O(Nlog^2N)$[11] where N is the number of nodes in the network.
- 4 communities found: 6017, 5432, 2, and 2. Both the communities with size 2 have an edge between them.
- Average rating of all the apps in both the communities is 4.19 and 4.11.
- Most of the paid apps are in community c_1(size 6017).

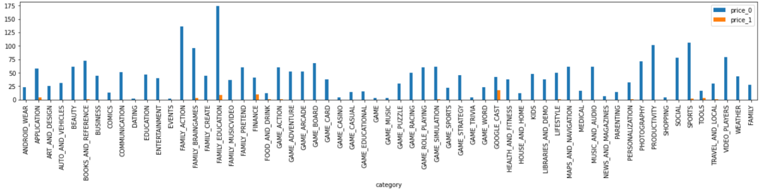

Figure: CNM community wise average price per category

[11] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. "Finding community structure in very large networks". In: *Phys. Rev. E* 70 (6 2004), p. 066111. DOI: 10.1103/PhysRevE.70.066111. URL: https://link.aps.org/doi/10.1103/PhysRevE.70.066111.

[12] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. "Exploring Network Structure, Dynamics, and Function using NetworkX". In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, 2008, pp. 11 –15.

Louvain Method

- Time complexity of $O(L)$[13] where L is the number of links.
- 3 communities found: 4329, 6114, 1010.
- 7 apps in c_3 out of 254 of the Finance category out of which 3 are paid (₹470, ₹700,₹700).
- The most expensive app in the whole dataset is ₹920.
- In Tools, there are only 6 apps in c_3 out of 179. 2 apps are paid.
- In both, the number of apps is less whereas their prices are high.

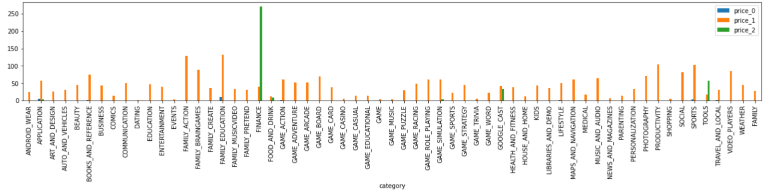

Figure: Louvain community wise average rating per category

[13] Vincent D Blondel et al. “Fast unfolding of communities in large networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/p10008. URL: http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

Infomap Community Detection

- The algorithm runs in $O(NlogN)$ with flow optimization[14].
- 3 communities found: 10531, 920, 2.
- Again the prices of c_1 (size 10531) are higher than c_2, with the same exceptions as for Louvain i.e. FINANCE and TOOLS.

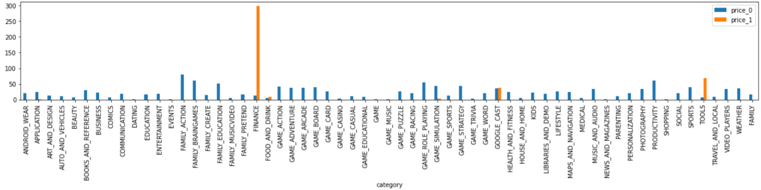

Figure: Infomap community wise average price per category

[14] Martin Rosvall and Carl T. Bergstrom. "Maps of random walks on complex networks reveal community structure". In: *Proceedings of the National Academy of Sciences* 105.4 (2008), pp. 1118–1123. ISSN: 0027-8424. DOI: 10.1073/pnas.0706851105. eprint: https://www.pnas.org/content/105/4/1118.full.pdf. URL: https://www.pnas.org/content/105/4/1118.

Outline

1. Introduction
2. Related Works
3. Dataset
4. Elementary Analysis of Original Network
5. Elementary Analysis of Sampled Network
6. Community Detection
7. **Survey**
8. Conclusion

Community Sizes

The size of the largest community predicted by CNM and Louvain is nearly 50% whereas for Infomap it is nearly 87% of the entire network and then drops quickly.

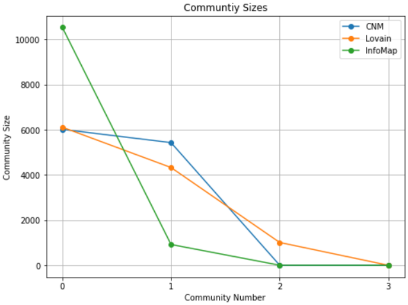

Figure: Community Sizes

Average Rating

- The average rating for community 2 is generally higher than other communities for all 3 algorithms.
- The rating of the largest community approximately coincides at 4.17 (4.19 for both CNM and Louvain, 4.15 for Infomap).
- This appears to indicate that a score of about 4.17 is preferable.

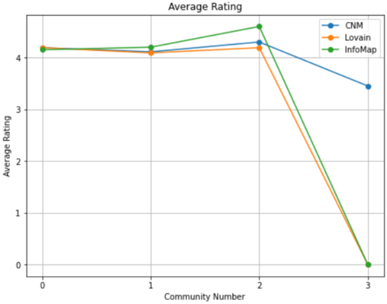

Figure: Average Rating

Average Price

- The average price of the largest community is much higher than other.
- For Louvain, the average price of community 3 is relatively high than CNM and Infomap and is comparable to the price of community 2 for CNM communities and then suddenly drops.

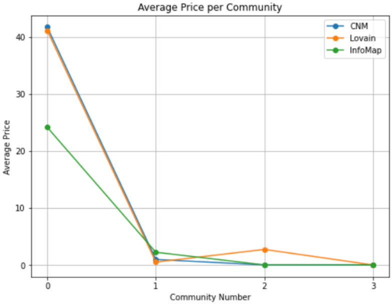

Figure: Average Rating

Average Reviews

- The number of reviews agrees at the largest community with approximately 869 reviews.
- Spike in communities 2 and 3 for Infomap and Louvain respectively.

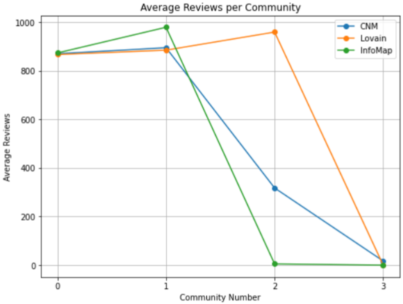

Figure: Average Reviews

Summary

- From the above analysis, it is clear that communities are not formed just based on the app's category but are a combination of various other factors including price and reviews.
- It seems that an app with a rating of near 4.17 and 870 reviews is more likely to get downloaded.
- Most of the paid apps have with a price near 40 are likely to get purchased.
- These findings can be used by Google play store to develop smarter recommendation engines or by developers to effectively advertise their application.

Outline

1. Introduction
2. Related Works
3. Dataset
4. Elementary Analysis of Original Network
5. Elementary Analysis of Sampled Network
6. Community Detection
7. Survey
8. Conclusion

Conclusion

- We explored the network formed by the apps of the google play store.
- We computed various centrality measures, checked scale-free nature, and tested a variety of network community detection algorithms.
- Able to identify many interesting communities with unique traits.
- The community detection algorithms we used were CNM, Louvain, and Infomap. Although the time complexities of these algorithms are similar, CNM took nearly 26 minutes to run whereas Louvain and Infomap were able to predict communities in nearly 5 minutes (tested on Google Colab).

References I

- https://zeenews.india.com/apps/mitron-app-suspended-from-google-play-store-2287704.html.

- Ashish Aggarwal. *Graph Sampling Package*. https://github.com/Ashish7129/Graph_Sampling.

- Vincent D Blondel et al. “Fast unfolding of communities in large networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/p10008. URL: http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

- Aaron Clauset, M. E. J. Newman, and Cristopher Moore. “Finding community structure in very large networks”. In: *Phys. Rev. E* 70 (6 2004), p. 066111. DOI: 10.1103/PhysRevE.70.066111. URL: https://link.aps.org/doi/10.1103/PhysRevE.70.066111.

- Google. *Google Play Store*. https://play.google.com/store. 2008.

References II

- Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. “Exploring Network Structure, Dynamics, and Function using NetworkX”. In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, 2008, pp. 11 –15.

- JoMingyu. *Google-Play-Scraper*. https://github.com/JoMingyu/google-play-scraper.

- Jure Leskovec and Christos Faloutsos. “Sampling from Large Graphs”. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’06. Philadelphia, PA, USA: Association for Computing Machinery, 2006, 631–636. ISBN: 1595933395. DOI: 10.1145/1150402.1150479. URL: https://doi.org/10.1145/1150402.1150479.

References III

- Julian McAuley, Rahul Pandey, and Jure Leskovec. “Inferring Networks of Substitutable and Complementary Products”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’15. Sydney, NSW, Australia: Association for Computing Machinery, 2015, 785–794. ISBN: 9781450336642. DOI: 10.1145/2783258.2783381. URL: https://doi.org/10.1145/2783258.2783381.

- *Number of free apps on Google play store*. 2021. URL: https://www.statista.com/.

- Facundo Olano. *google-play-scraper*. https://github.com/facundoolano/google-play-scraper. 2019.

References IV

- Martin Rosvall and Carl T. Bergstrom. “Maps of random walks on complex networks reveal community structure”. In: *Proceedings of the National Academy of Sciences* 105.4 (2008), pp. 1118–1123. ISSN: 0027-8424. DOI: 10.1073/pnas.0706851105. eprint: https://www.pnas.org/content/105/4/1118.full.pdf. URL: https://www.pnas.org/content/105/4/1118.

- Shihui Song and Jason Zhao. “Survey of Graph Clustering Algorithms Using Amazon Reviews”. In: ().