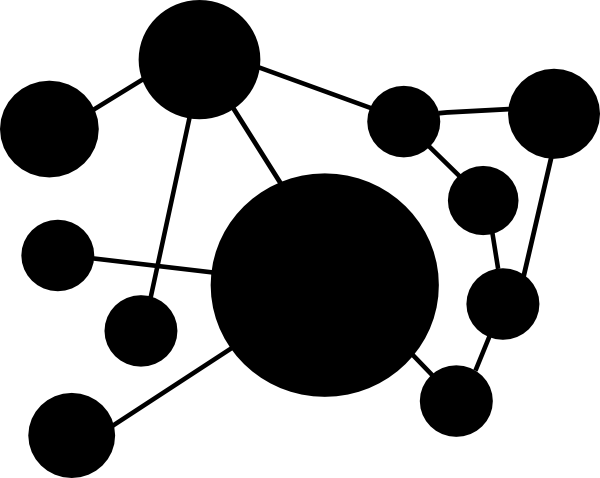


Modelling a Community:

Network Software and parkrun Data





|  |  |
| --- | --- |
| Student Name (Student Number) | Eoin Carroll (16202781) |
| Module Code | UCD MIS40550 |
| Module Title | Network Software Modelling |
| Assessment Title | Network Software Modelling Assignment 2 |
| Module Co-ordinator | Dr James McDermott |
| Revision | 0 |
| Date Submitted | 23/04/2017 |

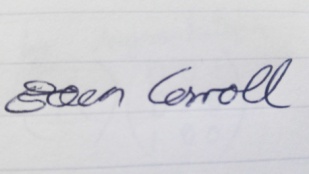
# Plagiarism

The unacknowledged inclusion of another person’s writings or ideas or works, in any formally presented work (including essays, examinations, projects, laboratory reports or presentations). The penalties associated with plagiarism designed to impose sanctions that reflect the seriousness of University’s commitment to academic integrity. Ensure that you have read the University’s Briefing for Students on Academic Integrity and Plagiarism and the UCD Plagiarism Statement, Plagiarism Policy and Procedures, (<http://www.ucd.ie/registrar/>).

# Declaration of Authorship

I declare that all material in this assessment is my own work except where there is clear acknowledgement and appropriate reference to the work of others.

# Signatures

Signed:  Date: 23/04/2017

# Abstract

Contents

[Introduction 4](#_Toc480321881)

[Graph Generation 4](#_Toc480321882)

[Web Scraping 4](#_Toc480321883)

[Constructing Graphs 4](#_Toc480321884)

[Combining Graphs 4](#_Toc480321885)

[Graph Properties 4](#_Toc480321886)

[Week 1 4](#_Toc480321887)

[Week 1&2 4](#_Toc480321888)

[Final Graph 4](#_Toc480321889)

[Information Flow 4](#_Toc480321890)

[Conclusion 4](#_Toc480321891)

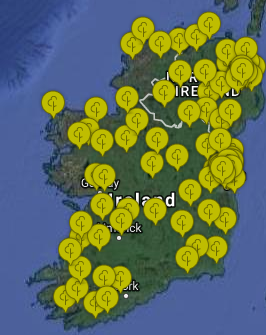
[References 4](#_Toc480321892)

[Appendix 4](#_Toc480321893)

# Introduction

Parkrun is a recent social phenomenon that has spread across 14 countries since 2004. It is a weekly timed 5km fun run on Saturday mornings, is free to attend and managed entirely by volunteers. Parkrun came to Ireland in 2012 and events now take place concurrently in 61 locations across the country (parkrun.ie, 2017). The result of this phenomenon is a rapidly growing community of runners and walkers.

In this project we will model the social network from one of these locations and examine the theoretical flow of information in the form of a message spreading through the community. Fortunately, Parkrun places event results online for us to explore and utilise. The scenario proposed is that a decision is made to cancel an event. This message is passed from person to person based on a probability that they know each other well enough to have each other’s phone number.

The event location of Marlay Park in Dublin has been selected as it is the most popular event in Ireland and will provide enough data for the study. We will examine the information flow in the following six cases:

1. Social network of the first event in Marlay park
2. Social network of the first 10 events (sample network)
3. Social network of all events to date (full network, 204 events)
4. Erdős–Rényi model with 100 nodes
5. Erdős–Rényi model with 1000 nodes
6. Erdős–Rényi model with 5000 nodes

# Graph Generation

One of the reasons parkrun was selected for the project is the availability of data. The results for each event online include each person’s finishing position, time, name and associate running club if available. Furthermore, parkrun has been experimenting with an API to allow direct access to the vast amount of data available. All data files are stored in a cloud hosting service linked in the appendix.

## Web Scraping

Unfortunately the API system was offline for the duration of the project and so alternative means of acquiring the dataset were explore. The next option explored was to sign up as a research partner which parkrun encourages however this was not practical due to time constraints. Next, the Python library Beautiful Soup was implemented to web scrape each results page for Marlay park. This was an interesting learning but the parkun website politely blocked web scraping. The final option was to manually copy data from each result into locally stored files and while this was not ideal, it sufficed for the project. Each weekly result file is saved in the data folder with the name (race week).xlsx.

Table 1 - Sample Header from dataset (simplified)

|  |  |  |  |
| --- | --- | --- | --- |
| Pos | parkrunner | Time | Club |
| 1 | Name 1 | 16:26 | Raheny Shamrock AC |
| 2 | Name 2 | 17:22 |  |
| 3 | Name 3 | 17:31 | Dundrum South Dublin Athletic Club |
| 4 | Name 4 | 17:36 | Activ Multisport |

## Constructing Graphs

An undirected, weighted graph was selected to represent the social network. Each node represents a unique name in the dataset, each edge is a connection between two people and self-loops were not applicable. The weight of each edge will represent the probability that these two people know each other well enough to have shared phone numbers.

Once the raw datasets had been saved from the results website, they could be processed into networks. An algorithm was constructed to read each name from the raw data and create nodes and weighted edges. This was completed individually for each race result set. The following attributes were used to assign probabilities of connections:

1. Two people attended the same event (everyone in each dataset)
2. Two people were in the same running club
3. Two people have the same second name
4. Two people finished with a similar finishing time
5. Two people finished with a similar position

Within the similar time and position parameters, same name and club were given more weighting to represent an increased probability. These parameters were tested with different inputs based mainly on domain knowledge. The resulting graphs were examined critically and the algorithm adjusted until a reasonable output was settled on. Note that this method allowed a small but present probability of all participants knowing each other in a given week. Parallel processing in python was implemented to speed up the network generation and the resulting network for each of the datasets was saved as a “edgelist” .csv file. Again, these data files can be seen at the attached link.

The next stage in the project was to merge the weekly networks as generate above, into a sample of the first 10 weeks and into a full dataset. All duplicate edge weights were summed together and included in the final networks. The graph properties section of this report includes an insight into the networks structure.

## Erdos-Renyi random graph model (Multiple Sizes)

# Graph Properties

Propose that anyone with a connection greater than 0.5 scaled knows each other.

Theoretical properties of Parkrun - all connected in each race. Possibility but unlikely that race would not have overlapping runners

Each race could be seen as a cluster

## Week 1

## Week 1 to 10

## Week 1 to 204

Full network – Marlay parkrun in it’s current state

## Graph Degree Plots

Degree Distribution (unscaled and log) for 5 graphs (First race will all be connected)

Degree Distribution with >0.5 for all 6 graphs

# Information Flow

Information passes based on scaled weight as a probability.

# Discussion

Note on run times - Run times of large graphs was significant. Could have made better . . .

Too much emphasis put on data generation – could have used random graphs instead of actual to save time. Arguably would have taken away from the authenticity of the project?

Added all edges together and summed weights if the edges appeared more than once. Combine time was under 10mins, Ram intensive. Final file size

## Further Research

A more thorough study would merge the networks from other races/locations

Once the API is back online, this could be easily done. http://www.parkrun.com/api/

Adjust graph generation parameters – is there a point of inflection? Complete a survey or something and recalibrate to ensure the results are accurate.

Duplicate names not handles very well - just skip name. Skip unknowns

Run time was CPU intensive and took 3 hours. Most likely ways to speed up.

# Conclusion

# References

Clipart: http://www.clker.com/cliparts/G/G/F/Y/Y/U/network-md.png

parkrun.ie accessed 2017

Parkrun Logo: http://www.parkrun.com/brand/

Stackoverflow – I mostly google things

# Appendix

Data location:

https://drive.google.com/drive/folders/0B9kelMwrpRsROF9UZFY5SzZZUzg

Include PC spec overclocked