Master Thesis

# Gaining Customer Insights using Machine Learning on Graphs

University of Basel

Author:
Michael von Siebenthal

Supervisor:
Prof. Dr. Dietmar Maringer

May 4, 2021

# Abstract

Abstract goes here

# Declaration

"I hereby declare - that I have written this master thesis without any help from others and without the use of documents and aids other than those stated in the references, - that I have mentioned all the sources used and that I have cited them correctly according to the established academic citation rules, - that the topic or part of it are not already the object of any work or examination of another course unless explicitly stated,- that I am aware of the consequences of plagiarism at the Business and Economics Faculty of University of Basel."

Michael von Siebenthal, Martikel-Nr.: 2015-256-837, Date: May 4, 2021

# Contents

# List of Figures

# List of Tables

# Introduction

The aim of this thesis is to explore the relatively new field of machine learning on graphs for the purpose of gaining customer insights. Graph machine learning is the current frontier in machine learning and has vast applications in many areas as recently shown by the success of AlphaFold (Senior et al. 2020). AlphaFold made a breakthrough for predicting protein structures where they made use of the observation that a folded protein can be considered as a spatial graph (AlphaFold 2020). In addition, there are a vast range of applications for graphs in the fields of natural science, social science and many more as shown by the excellent overview given by Zhou et al. (2020). In principle, graphs are useful whenever one wants to model interactions or relationships.

From a business & economics perspective, graphs are particularly interesting if one wants to for instance model the interactions between institutions. An example for this is a study published by Schweitzer et al. (2009) which created a graph showing the interdependencies of international banks as a network, see figure 1.1. This is a useful representation of interdependencies and is an important basis for making systems more robust.

Another interesting application of graphs for business & economics are social interactions. While there are many different types of social interactions of interest, social interactions for marketing purposes have been among the most popular. Indeed, this is one of the main areas where social networks such as Facebook or search providers such as Google make their revenue by selling advertising (Facebook 2021, Alphabet 2021).
Both Facebook and Google have the advantage, that their businesses naturally capture relational or more generally network data which can be represented as graphs. Most researchers or companies however do not have access to such data. Companies for instance may have access to large amounts of customer data, however they typically would not have access to relational information (e.g. which client is connected with which other clients). The same is true for researchers, where social scientists

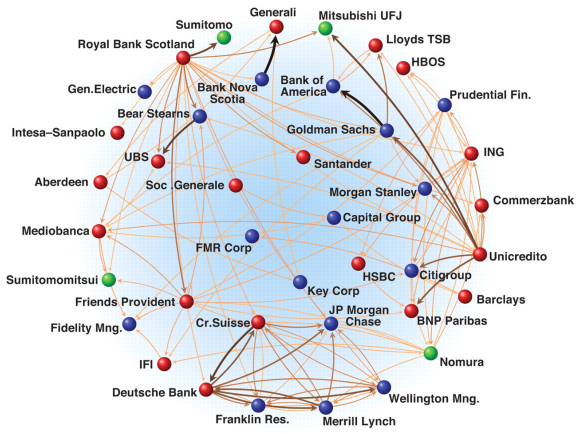


Figure 1.1: Bank Network
(Schweitzer et al. 2009, p. 424)

often collect data via anonymous surveys. This is an issue in terms of data access.

This leads us to the two main aspect of this master thesis which consists of:

1. Methodology
2. Application

Given the difficult access to graph data, this thesis will explore whether (semi-) synthetic graphs can be generated using cross-sectional data gathered from surveys. The aim is not only to generate graphs but to then test whether the resulting graph can be used for meaningful machine learning tasks. In order to test this, cross-sectional data of banking clients will be gathered. The aim in terms of application will be to perform an investor classification task. This application area is chosen due to the author's experience in this field having worked as a client adviser at a bank for over 10 years. One could however choose almost any customer classification area, meaning that the application can be chosen arbitrarily within reason.

It would of course be best, if one would have access to real graph data. If graph data had been available for a unique application, this would have been preferred for this thesis. The absence of such available graph data and the difficulty of generating graph data sparked the interest for exploring synthetic graph generation for subsequent machine learning tasks. If this application proved to be successful, this could provide an alternative approach for analyzing cross-sectional data. Given the richer amount of information.

## 1.1 Overview Application

Retail banking is an area in which client advisers typically serve several hundred if not thousands of clients. In addition, advisers typically work in teams which makes personalized advice virtually impossible. For this reason, it is impossible for an adviser to ensure that she provides offers to her clients which are in line with their interests. From personal experience of the author, having worked as an adviser for over 10 years, this can lead to unsatisfactory outcomes when contacting clients for services or products that they are not interested in. This is largely due to inadequate client selection practices. In a bank setting, clients for marketing campaigns, such as promoting investment products, are typically selected based on whether they have available liquidity for investment and/or whether a client has already invested in similar products. While this selection makes sense to a certain extent, it falls short in that there remain often too many clients in the selection, which are not interested in the offer of a specific marketing campaign or potentially interested clients are falsely excluded in the pre-selection process.

This is an area where classifying clients according to their interests could be of great value and where machine learning methods could be of particular use. It would in addition improve the service quality rendered to clients, prevent unnecessary and perhaps annoying marketing calls and improve the efficiency of a marketing campaigns.

## 1.2 Literature Review

There is a lot of published research regarding bank client classification using a myriad of methods. Classifications tasks are performed in areas such as credit scoring, anti-money laundering (AML) compliance or marketing purposes among others (Sukharev et al. 2020, Weber et al. 2018, Moro et al. 2014).

CONTINUE ANOTHER TIME

# Theory

This chapter will cover the necessary theoretical background for this master thesis and consists of following parts:

1. Graph Theory
2. Machine Learning on Graphs
3. Graph Generation

## 2.1 Graph Theory

This section provides a brief introduction to graph theory, with a focus on the relevant aspects for this master thesis. The theory presented is primarily taken from the book "Networks: An Introduction" by Mark Newman (2010).

ADD AS BACKGROUND graph theory is an old field of mathematics and can be traced back to Leonhard Euler and the famous "Knigsberg Bridge Problem" (Euler 1736), it has experienced a recent revival in machine learning. As it is a relatively novel field in machine learning, it provides for a fruitful ground for testing methods and applications.

Graphs are special data structures as shown in Figure 2.1. The terms graph and network are used inerchangeably for the purpose of this master thesis. Typically, the term graph is used more commonly when referring to mathematical analysis and the term network is more commonly used for data science purposes.
The graph shown in Figure 2.1 corresponds to an undirected graphs in which the connections between the vertexes are mutual. In a directed graph for instance, vertex A could be connected to vertex B, however vertex B need not be connected to vertex A. For the purpose of this thesis, only undirected graphs are considered. Vertexes are commonly reffered to as nodes and the terms will be used interchangeably. Edges refer to the connections between the vertices. Edges are often also referred to

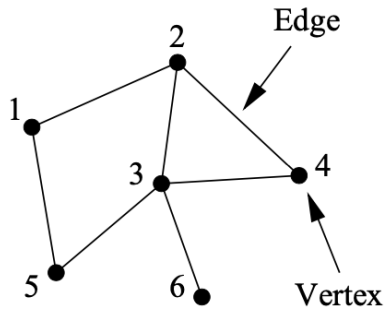


Figure 2.1: Example of a Graph
(Newman 2010, p. 111)

as links and the terms will be used interchangeably as well. Graphs may have additional elements such as multi-edges or self-edges. These are however not relevant for the purpose of this master thesis.

In terms of mathematical notation, graphs are typically defined as follows:

$$G(V, E) \tag{2.1}$$

$G$ refers to the graph as an output. $V$ refers to the set of vertices present in the graph and $E$ refers to edges present between the vertices.

### 2.1.1 Adjacency Matrix

The adjacency matrix $A$ is defined as a $n \times n$ matrix, where $n$ refers to the number of vertices present in the graph. Each vertex is therefore recorded by a column and a row in the adjacency matrix. The elements in the adjacency matrix are further typically defined as follows:

$$A_{ij} = \begin{cases} 1, & \text{if vertex } i \text{ and } j \text{ are connected by an edge} \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

For illustration, the adjacency matrix of the graph shown in Figure 2.1 is shown as follows:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

As one can see, if vertex $i$ and $j$ are connected, this is recorded with 1 and 0 otherwise. Note, that all the elements on the $diag(A)$ are equal to 0. This is because self-edges are excluded. As this is an undirected network, the adjacency matrix is symmetric. There are many additonal aspects one could mention with regard to the adjacency matrix, they are however not relevant for this thesis.

### 2.1.2 Degree Measures

An important measure for graphs are the degrees $k$ of the vertices, the number of edges connected to a vertex. The degrees of vertex $i$ can be formulated as (Newman 2010, p.133):

$$k_i = \sum_{j=1}^{n} A_{ij} \tag{2.3}$$

For an undirected graph, edges have two ends. This is due to the fact that vertices connected by an edge are mutually connected. In terms of the sum of the degrees of all vertices, we can therefore write for a graph with $m$ edges (Newman 2010, p.133):

$$2m = \sum_{i=1}^{n} k_i \tag{2.4}$$

The sum of all degrees is therefore just the number of edges $m$ multiplied by 2. In terms of statistical measures, the mean degree $c$ of a vertex is defined as follows (Newman 2010, p.134):

$$c = \frac{1}{n} \sum_{i=1}^{n} k_i = \frac{2m}{n} \tag{2.5}$$

In order to define the connectance or density of a graph, we must first observe, that the maximum number of edges is given by (Newman 2010, p.134):

$$\binom{n}{2} = \frac{1}{2} n(n-1) \tag{2.6}$$

The density $\rho$ can therefore be written as (Newman 2010, p.134):

$$\rho = \frac{m}{\binom{n}{2}} = \frac{2m}{n(n-1)} = \frac{c}{n-1} \tag{2.7}$$

Note, that the density $\rho$ lies strictly between $0 \leqslant \rho \leqslant 1$. In addition, for sufficently large graphs, one can approximate $\rho = \frac{c}{n}$.

### 2.1.3 Eigenvector Centrality

The degrees of a vertex shown in the previous section already correspond to the simplest form of centrality measures. The issue with this measure however is, that the every neighbor of vertex $i$ are valued the same. This is a problem, as not all neigbors are of equalt importance due to:

1. Number of neighbors
2. Importance of neighbor
3. both

There are many different alternative centrality measures which can consider the factors listed above such as eigenvector ecntrality, Katz centrality or PageRank (Katz 1953, Page et al. 1999, Landau 1895, Newman 2010). As we are only dealing with simple undirected graphs, eigenvector centrality will suffice, where the other mentioned methods are adaptations to the eigenvector centrality.

Eigenvector centrality gives all vertices a score proportional to the sum of the scores of its neighbors. This is a procedure in which typically the initial centrality $x_i$ of vertex $i$ is guessed to be $1 \; \forall \; i$. This can be used to calculate the centralities of the neighbors of $i$ which is denoted as $x'_i$. We can thus write (Newman 2010, p. 169):

$$x'_i = \sum_j A_{ij} x_j \tag{2.8}$$

In matrix form:

$$x' = Ax \tag{2.9}$$

This process is repeated $t$ times to provide better estimates (Newman 2010, p. 170):

$$x(t) = A^t x(0) \tag{2.10}$$

Where $x(0)$ is a linear combination of (Newman 2010, p. 170):

$$x(0) = \sum_i c_i v_i \tag{2.11}$$

$v_i$ correspond to the eigenvectors of the adjacency matrix $A$ and $c_i$ corresponds to an appropriately chosen constant. Therefore we can write (Newman 2010, p. 170):

$$x(t) = A^t \sum_i c_i v_i = \sum_i c_i k_i^t v_i = k_i^t \sum_i c_i \left[\frac{k_i}{k_1}\right]^t v_i \tag{2.12}$$

In the above equation, $k_i$ correspond to the eigenvalues of $A$, the adjacency matrix. $k_1$ corresponds to the largest eigenvalue of $A$. As $\frac{k_i}{k_1} < 1, \ \forall \ i \neq 1$ , the term is decaying as $t \to \infty$. The centralities $x$ can therefore be written as fulfilling following condition (Newman 2010, p. 170):

$$Ax = k_1 x \tag{2.13}$$

Lastly, the eigenvector centrality is defined as (Newman 2010, p. 170):

$$x_i = k_1^{-1} \sum_j A_{ij} x_j \tag{2.14}$$

### 2.1.4 Closeness Centrality

Closeness centrality, $C_i$, is defined as the average distance from a vertex to the other vertices. This centrality measure is defined as follows (Newman 2010, p. 182):

$$C_i = \frac{1}{l_i} = \frac{n}{\sum_j d_{ij}} \tag{2.15}$$

In this measure, central vertices exhibit high closeness centrality and are therefore closer connected to other vertices compared to vertices with low closeness centrality. $l_i$ refers to the average of the geodesic distances $d_{ij}$ of vertex $i$.

### 2.1.5 Betweenness Centrality

This centrality measures to which extent a vertex lies on paths between other vertices. For instance, a bottle neck vertex would exhibit a large betweenness centrality as many, if not all nodes must pass through it. More formally, betweenness centrality, $x_i$, is defined as (Newman 2010, p. 187):

$$x_i = \sum_{st} \frac{\eta_{st}^i}{g_{st}} \tag{2.16}$$

In the above equation, $\eta_{st}^i$ refers to the number of geodesic paths from $s$ to $t$ which

pass through vertex $i$. Further, $g_{st}$ is defined as the number of geodesic paths between vertex $s$ and $t$.

In order to allow for better comparison of betweenness centrality, it is often standardized by the number of connected vertex pairs $s$ and $t$ denoted as $\eta^2$. The betweenness centrality can therefore be expressed as (Newman 2010, p.190):

$$x_i = \frac{1}{\eta^2} \sum_{st} \frac{\eta^i_{st}}{g_{st}} \tag{2.17}$$

With this measure, the betweenness centrality is within the range $0 \leqslant 1$

## 2.2 Machine Learning on Graphs

WHY GRAPH ML Graph structures are significantly different compared to traditional types of data usually used for machine learning tasks or regression analyses. Graph structures are special in that the data points in a graph have connections with each other. A practical example for this are social networks. In a social network, the profiles of "Peter" and "Paul" might be connected because Peter and Paul are friends. This aspect is unique to graph or network data and provides both additional information and additional complexity to graph data not found in other data types. An introduction to graph theory will be given in chapter 2.

## 2.3 Graph Generation

It is very difficult to collect graph data, as one cannot capture links between observations by traditional data collection means such as surveys. In this regard, banks have a unique advantage (similar to social networks) in that they can capture relationships by analyzing their clients cash flows. In today's world most people make most if not all of their payments electronically and usually pay by card. This provides banks with a large amount of information such as spending behavior. Payments are nothing other than links in a graph theoretical setting. Unfortunately, access to bank client data is rarely granted due to privacy or bank secrecy laws.

This is a problem for master thesis for which no perfect solution exists. The search for alternatives lead to graph generation methods. Among the many graph generation algorithms researched, the Multiplicative Attribute Graph (MAG) model appears to provide a feasible solution (Kim & Leskovec 2012). This model creates probabilistic links between observations using link-affinity matrices. This section

will provide an introduction to the MAG model.

# Bibliography


Alphabet (2021), 'Alphabet investor relations'. (accessed: 04.05.2021).
**URL:** *https://abc.xyz/investor/*

AlphaFold (2020), 'Alphafold: a solution to a 50-year-old grand challenge in biology'. (accessed: 05.05.2021).
**URL:** *https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology*

Euler, L. (1736), 'Solutio problematis ad geometriam situs pertinentis', *Commentarii academiae scientiarum Petropolitanae* **8**, 128–140.

Facebook (2021), 'Facebook investor relations'. (accessed: 04.05.2021).
**URL:** *https://investor.fb.com/investor-events/default.aspx*

Katz, L. (1953), 'A new status index derived from sociometric analysis', *Psychometrika* **18**(1), 39–43.
**URL:** *https://doi.org/10.1007/BF02289026*

Kim, M. & Leskovec, J. (2012), 'Multiplicative attribute graph model of real-world networks', *Internet mathematics* **8**(1-2), 113–160.
**URL:** *https://doi.org/10.1080/15427951.2012.625257*

Landau, E. (1895), 'Zur relativen wertbemessung der turnierresultate', *Deutsches Wochenschach* **11**, 366–369.

Moro, S., Cortez, P. & Rita, P. (2014), 'A data-driven approach to predict the success of bank telemarketing', *Decision Support Systems* **62**, 22–31.
**URL:** *https://doi.org/10.1016/j.dss.2014.03.001*

Newman, M. (2010), *Networks: An Introduction*, Oxford University Press, Inc.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999), The pagerank citation ranking: Bringing order to the web., Technical report, Stanford InfoLab.
**URL:** *http://ilpubs.stanford.edu:8090/422/*

Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A. & White, D. R. (2009), 'Economic networks: The new challenges', *science* **325**(5939), 422–425.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A. et al. (2020), 'Improved protein structure prediction using potentials from deep learning', *Nature* **577**(7792), 706–710.
**URL:** *https://doi.org/10.1038/s41586-019-1923-7*

Sukharev, I., Shumovskaia, V., Fedyanin, K., Panov, M. & Berestnev, D. (2020), 'Ews-gcn: Edge weight-shared graph convolutional network for transactional banking data', *arXiv preprint arXiv:2009.14588* .
**URL:** *https://arxiv.org/abs/2009.14588*

Weber, M., Chen, J., Suzumura, T., Pareja, A., Ma, T., Kanezashi, H., Kaler, T., Leiserson, C. E. & Schardl, T. B. (2018), 'Scalable graph learning for anti-money laundering: A first look', *arXiv preprint arXiv:1812.00076* .
**URL:** *https://arxiv.org/abs/1812.00076*

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C. & Sun, M. (2020), 'Graph neural networks: A review of methods and applications', *AI Open* **1**, 57–81.
**URL:** *https://doi.org/10.1016/j.aiopen.2021.01.001*