Changing Perspectives: Using Graph Metrics to Predict Purchase Probabilities

Article *in* Expert Systems with Applications · October 2017

DOI: 10.1016/j.eswa.2017.10.046

CITATIONS
15

READS
1,521

4 authors:

Annika Baumann
Universität Potsdam
39 PUBLICATIONS **420** CITATIONS
SEE PROFILE

Johannes Haupt
Humboldt-Universität zu Berlin
12 PUBLICATIONS **91** CITATIONS
SEE PROFILE

Fabian Gebert
WirMarkt
5 PUBLICATIONS **57** CITATIONS
SEE PROFILE

Stefan Lessmann
Humboldt-Universität zu Berlin
141 PUBLICATIONS **2,986** CITATIONS
SEE PROFILE

Changing Perspectives: Using Graph Metrics to Predict Purchase Probabilities

Annika Baumann[a,*], Johannes Haupt[b], Fabian Gebert[c], Stefan Lessmann[d]

[a] Chair of Information Systems, Humboldt University of Berlin, Spandauer Straße 1, 10178 Berlin, Germany, e-mail address: annika.baumann@wiwi.hu-berlin.de

[b] Chair of Information Systems, Humboldt University of Berlin, Spandauer Straße 1, 10178 Berlin, Germany, e-mail address: johannes.haupt@hu-berlin.de

[c] Akanoo GmbH, Mittelweg 121, 20148 Hamburg, Germany, e-mail address: fabian@akanoo.com

[d] Chair of Information Systems, Humboldt University of Berlin, Spandauer Straße 1, 10178 Berlin, Germany, e-mail address: stefan.lessmann@hu-berlin

[*]Corresponding author: annika.baumann@wiwi.hu-berlin.de, +49 30 2093-5740

Abstract

The prediction of online user behavior (next clicks, repeat visits, purchases, etc.) is a well-studied subject in research. Prediction models typically rely on clickstream data that is captured during the visit of a website and embodies user agent-, path-, time- and basket-related information. The aim of this paper is to propose an alternative approach to extract auxiliary information from the website navigation graph of individual users and to test the predictive power of this information. Using two real-world large datasets of online retailers we develop an approach to construct within session graphs from clickstream data and demonstrate the relevance of corresponding graph metrics to predict purchases.

Keywords: Predictive analytics, Clickstream data, User graph, Graph metrics.

1. Introduction

The e-commerce sector is responsible for a substantial fraction of firm revenues. Annual turnover was 1,336 billion US dollars in 2014 and is predicted to have reached 2,050 billion US dollars in 2016 (Statista, 2016a). However, given that growth rates are expected to decline in the future (Statista, 2016b), e-commerce shops need to find ways to defend market shares in an increasingly competitive environment. One strategy is to increase purchasing amounts and/or frequencies of existing customers. Important determinants of (re-)purchase intention in online shopping are trust, service quality (Hong & Kim, 2012) and user satisfaction during the online shopping process (Lee et al., 2009). To offer a richer user experience and increase visitors' (re-)purchase intentions, understanding customer online behavior is crucial (e.g., Pai et al., 2014). To gain such insight and to anticipate user actions, the analysis of clickstream data has been widely adopted in the literature (e.g., Van den Poel & Buckinx, 2005; Park and Park, 2016).

However, previous work in the field has not examined the potential of graph theory to gather auxiliary information from clickstream data and increase the accuracy of behavior prediction models. Graphs are a methodological approach originating from network theory. They consist of nodes and edges, which connect nodes. Graph-based approaches have been used in various fields and have been proven to be helpful for various tasks, for example to predict connections in the social networking context (He et al., 2015), to detect money laundering activities (Colladon & Remondi, 2017), for personalized recommendations (Shams & Haratizadeh, 2017) and for customer churn prevention (Óskarsdóttir et al., 2017). Given the success of graph-based predictors in these and other applications, the objective of our paper is to test their potential for online behavior prediction based upon clickstream data.

We contribute to literature as follows: First, we propose an approach to derive graphs from user sessions based on clickstream data. Second, we calculate graph metrics and examine their pairwise dependency in terms of correlation. Third, we assess how they perform as a means to predict customer behavior in online contexts.

The remainder of the paper is structured as follows. First, we give an overview on relevant literature to clarify the research gap the paper strives to close. Afterwards, we present our methodology and how we derive clickstream graphs in particular. We then summarize the resulting data, before presenting empirical results. Lastly, we summarize our findings.

2. Related Work

Much literature considers the use of clickstream data for customer online behavior prediction. Prediction targets range from conversions in purchase prediction (Van den Poel & Buckinx, 2005), whether visitors redeem incentives (Pai et al., 2014) or complete specific tasks such as putting an item into a basket (Sismeiro & Bucklin, 2004; Kalczynski et al., 2006), over navigational behavior

prediction (e.g., the next web path access; Montgomery et al., 2004) to classifying visitors into interest groups such as whether a user's site visiting intention is informational or transactional (Moe, 2003).

Table 1 summarizes related work, which we categorize according to the target of prediction into navigational behavior (NB), user classification (UC) and conversion (PC) prediction, where PC is the prevailing target in prior work (i.e., 23 out of 34 studies fall in this category).

Reference	Dependent Variable	Feature Category					
		Page	Time	Monetary	Page Interaction	Demo-graphics	Graph / Similarity
Anitha 2010	NB	x					
Antonellis et al. 2009	UC	x					
Banerjee & Ghosh 2001	UC	x	x				x
Berka & Labsky 2007	NB	x					
Byeon 2013	PC	x	x				x
Chan et al. 2014	PC	x			x	x	
Girija & Kavitha 2013	NB	x					
Iwanaga et al. 2016	PC	x	x				
Jiang et al. 2012	PC	x	x				
Kalczynski et al. 2006	PC	x					x
Lee et al. 2010	PC	x	x				
Lu et al. 2005	UC	x					
Moe 2003	UC	x	x				
Moe & Fader 2004	PC		x	x			
Moe et al. 2002	PC	x	x				
Montgomery et al. 2004	NB	x					
Gündüz & Özsu 2003	NB	x	x				x
Padmanabhan et al. 2006	PC / Revisit	x	x	x		x	
Pai et al. 2014	UC	x	x				
Panagiotelis et al. 2014	PC	x	x	x			
Park et al. 2008	UC	x					
Park & Park 2016	PC	x					
Pitman & Zanker 2010	PC	x			x		
Sarwar et al. 2015	PC	x	x	x			
Sato & Asahi 2012	PC (day)	x		x			
Senecal et al. 2014	UC	x	x		x		
Sismeiro & Bucklin 2004	PC	x	x		x		
Stange & Funk 2015	PC	x	x	x	x		
Suh et al. 2004	PC	x	x				
Van den Poel & Buckinx 2005	PC	x	x	x		x	
Vroomen et al. 2005	PC		x	x	x	x	
Wu et al. 2005	PC	x					
Zhao et al. 2016	PC	x	x	x			
Zheng et al. 2003	PC	x	x	x		x	

Table 1. Overview of feature categories used in research (NB: Navigational Behaviour, UC: User Classification, PC: Conversion).

Table 1 also shows the types of features (i.e., covariates) which the studies employ for predictive modelling. In particular, we categorize the features into six groups. All categories except demographics are based on clickstream data. The first three groups – time, page and monetary – draw inspiration from the well-known concept of recency, frequency and monetary value analysis (Zhang et al., 2015). Recency and frequency consist of aspects such as time on page and last website visit (*Time*), whereas monetary comprises historical purchase behavior derived from preceding clickstream sessions and current basket information (*Monetary*). Frequency refers to the path traversal and categories of pages visited, counting how often each page has been visited (*Page*). In addition, we consider behavior related variables (*Page Interaction*), such as basket interaction, click on page and scroll on page events to capture user-centered feature categories, which revolve around behavioral aspects besides the website path that a user traverses. The feature category *Demographics* consists of variables that capture user characteristics, such as gender and geographic-related information, which are not related to the website itself and thus not part of clickstream data. The last category captures studies which use graphs as a tool to derive features for their models used (*Graph / Similarity*).

Only four of the 34 studies, which base their analysis on clickstream, use a graph-based approach. In view of Table 1 it becomes evident that combining predictive modelling with graph-based features has been rare so far. Byeon (2013) generates a bi-partite graph (i.e. a graph with two different types of nodes) for each user session, where the nodes represent a specific webpage a user visits during her session and the category to which the webpage belongs, respectively. Each graph facilitates the calculation of summary statistics (i.e., density), which Byeon (2013) employs to predict whether a user session leads to a purchase using a logistic regression classifier. In comparison to classical clickstream features (e.g. total number of clicks, total visit time), the graph density feature provides encouraging results, suggesting that it is a good predictor of purchase intention.

Kalczynski et al. (2006) predict whether a specific website task has been completed successfully. To achieve this, they focus on navigational complexity. The authors construct an experimental setup where users are asked to browse a website to conduct an artificial purchase. The website data is based on five different datasets which they use to derive graphs from user journeys on a website. The authors then calculate a set of graph measures, some of which are based on specific website characteristics. Finally, they employ logistic regression to predict online task completion and conversion in particular.

Other approaches aiming at user classification do so via clustering using graph-based approaches to be able to build similarity graphs, connecting users which behave similarly on websites. Banerjee and Ghosh (2001) use clickstream data to create a similarity graph, which connects users who display similar website usage behavior. First, they select pair-wise user sessions and compare them in terms of path and time dimensions to derive a similarity score. They then construct a weighted graph with nodes representing users which are pairwise connected once the weight reaches a specific threshold. The weight represents the similarity between two users. The similarity graph serves as input to a graph-based clustering method to derive user groups.

A similar approach is applied by Gündüz & Özsu (2003) who construct a similarity graph to apply a graph-based clustering method. Their graph is based on path and time aspects associated with a user journey on a website. The aim of graph construction and clustering is to predict the next website request.

The review of related work suggests that a comprehensive study, which systematically assesses the predictive value of a broad set of graph metrics is lacking. Building on the work of Byeon (2013) and Kalyznski et al. (2015) to predict purchase intention/conversion, we contribute toward closing this research gap in that we i) develop a way to derive a graph from clickstream data, ii) consider a much richer set of graph metrics, iii) employ real-life data, and iv) use state-of-the-art prediction algorithms (random forest, gradient boosting machine) alongside logistic regression.

3. Methodology

The following sections explain our approaches to create clickstream graphs and derive corresponding graph metrics as input for predictive modeling.

3.1 Clickstream and Graph Construction

A clickstream is defined as the path which a website user traverses when visiting a number of websites (Bucklin et al., 2002) and consists of sessions each of which represent a single visit of a user on a website. Each session consists of an arbitrary number of page views, which are the webpages the user visits during a session. Specific behavior can be performed on a webpage such as click, scroll and basket events. Furthermore, single webpages are visited for a specific amount of time. So far, the representation of clickstream in the form of a graph has been established mainly for visualization purposes (e.g., Kitts et al., 2002). Using clickstream data, we construct a graph for each session of a user to be able to derive covariates for purchase prediction. In general, a graph $G = (V, E)$ consists of a set of nodes V which are pair-wise connected via edges E. The edges are either directed or undirected. Each graph can be represented as a $n \times n$ adjacency matrix where elements a_{ij} are set to one if node n_i and n_j are connected, and zero otherwise.

The user session graphs applied in this paper are constructed in the following way: Each node represents a specific website a user has visited during the session. For each page view, we create a new node if it does not already exist in the session graph (i.e., the user has not visited the page before). We connect two nodes with an edge (i.e., between two pages), if the user visits them successively. The edges are directed to capture the specific order in which webpages are visited. Due to incremental node insertion, the session graph grows successively during the users' journey on the website. This technique is known as "clipping at every click" (Van der Meer et al., 2000), meaning that, for every page view, we calculate a new graph and its underlying graph metrics to capture the user sessions' characteristics in an incremental manner. Figure 1 shows an example of this approach where a session of a user is represented as a graph structure which is updated at every page view, i.e., every webpage the user visits during her journey on the website. Furthermore, as an example the incremental calculation of a graph metric (i.e., average in-degree, which is the average number of edges converging to a node) is shown, which is re-calculated at each page view of a user.

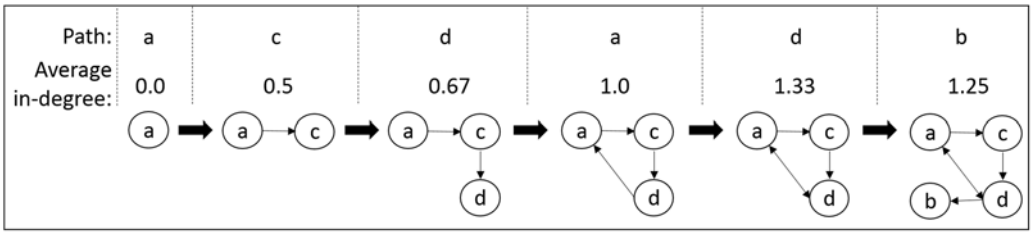

Figure 1. Example of a graph inference of a user session based on clickstream data.

Based upon Kalczynski et al. (2006), we assume that the structure of a graph represents user behavior, which motivates examining the potential of graph metrics to predict conversion. More specifically, a graph based on clickstream data grounds on the explicit user click behavior and captures the path traversal of a user on a website. This behavior is a direct result of the user's goal in visiting the website and changes observably with user intention. To illustrate this, exemplary session graphs for three different types of user behavior are shown in Figure 2.

The left graph shows a direct clickstream path, traversing from one page to another without returning behavior. This indicates a high degree of goal-orientation, which can be associated with both informational (the sought-after piece of information was found) or transactional (the desired product was bought) behavior. The middle graph illustrates a customer looking at various products of two types of product categories before deciding which is of further interest. This is a typical comparison behavior. The right-most graph depicts broad browsing behavior. It contains several central nodes

signifying, for example, the overview pages of search results to which the user returns after looking at a sequence of products. These and other graph structures are captured by the graph metrics we apply and reduce to input suitable for predictive modelling.

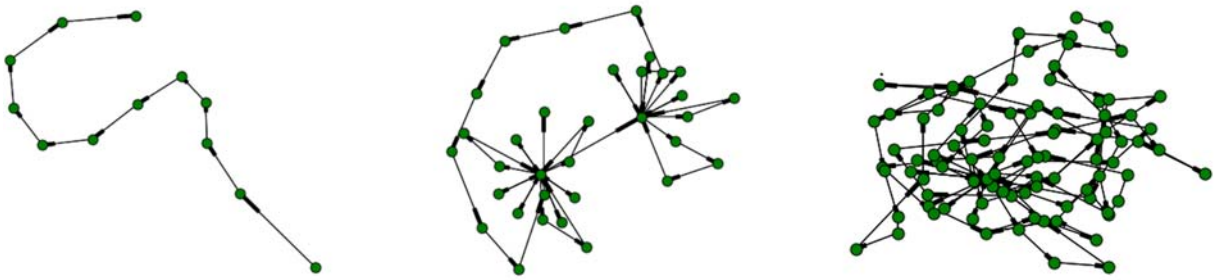

Figure 2. Graph visualizations of user sessions representing different types of user behavior.

3.2 Selected Graph Metrics

Several metrics have been established in graph theory to represent the characteristics of a graph. These metrics focus on a specific node, the n-hop neighborhood or the whole graph. Given our objective to clarify the relative predictiveness of different graph metrics in combination with the scarcity of prior work on graph-based online behavior prediction, it is not clear which graph metrics are the most informative. Therefore, we use the Python framework NetworkX (Hagberg et al., 2008) to create a large number of alternative metrics and compare their relative predictiveness empirically. To that end, we use a set of 23 graph measures in total (see Table 2). Metrics focusing on characteristics of single nodes are computed for all nodes in the graph and then averaged. We concentrate on structural, centrality- and distance-based metrics since they are able to describe the relative importance of graph elements in the network and the general structure of a graph. This in turn might be indicative of the users' intention behind the website visit.

Structural measures describe the general construction of a graph. The most basic concepts are the total number of nodes *N,* capturing the total number of unique webpages which a user visits, and the total number of edges *M,* being the number of directed and unique path traversals from one page to another. The number of circles and self-loops in a graph accounts for switching behavior of a user returning to a previously visited page or the same page, respectively. Related to these measures is flow hierarchy which states the proportion of edges not being part of a cycle. Transitivity indicates whether a user resolves around a specific subset of webpages such as switching between different products to choose from. Unlike transitivity, which focuses on the neighborhood of a node, density can be used as an additional measure of purposefulness of a session, since small values indicate a step-by-step list of pages without circularity or jump-backs hinting at more goal-oriented user behavior. Last, high values of the metric average node connectivity signal an interwoven connection between a considerable number of nodes and therefore a non-structured browsing behavior of users.

Distance measures relate to the broadness of a graph structure. The average shortest path length and the related metrics eccentricity, diameter, radius, center and periphery give an indication how diverse the user traverses through the website. Their intuitive interpretation is that for high values of, e.g., the average shortest path length, the user rather looks at unique webpages one after another without the occurrence of returning behavior. Small values signal returning behavior to formerly visited pages such as overview and search result pages.

Centrality measures describe the importance of nodes in terms of how central they are located in the graph structure. For example, a particular node can be described by its degree centrality which can be interpreted as the amount to which users return to a specific page from varying other pages. Betweenness centrality for both nodes and edges measures whether there are bridging elements in the network structure, such as specific overview pages which a user frequently returns to in order to access other webpages. Therefore, both concepts can be seen as examples for comparison behavior of

a user. Eigenvector, katz and pagerank centrality indicate whether there is a wide choice of disjoint paths, where high average values indicate an interwoven structure of several important nodes. The intuition of closeness centrality and closeness vitality is that both measures assume high values if a node is located central in the whole network. For example, this applies to specific webpages a user has visited several times during the whole session. This extends the notion that there may be a specific number of pages that are central in the clickstream.

The summary statistics for the graph metrics of both shops is provided in the appendix.

Category	Metric	Feature	Description
Structure (8 metrics)	Number of nodes	NumberNodes*	Total number of nodes in the graph.
	Number of edges	NumberEdges*	Total number of edges in the graph.
	Number of cycle	NumberCircles	Total number of circles in the graph.
	Number of self-loops	SelfLoops	Total number of self-loops in the graph.
	Flow hierarchy	FlowHierarchy	Proportion of edges not being part of a cycle.
	Transitivity	Transitivity	The number of triangles in the graph divided by the maximum possible number of triangles.
	Density	Density	The sparseness in terms of connectivity for the whole graph.
	Average node connectivity	NodeConnectivity*	Average number of nodes for each distinct node pair that must be removed from the network in order to disconnect them.
Distance (6 metrics)	Average shortest path length	ShortestPath*	The average of the shortest path length for all distinct node pairs in the graph.
	Average eccentricity	Eccentricity*	Average of the longest shortest path for each single node in the graph.
	Diameter	Diameter*	The maximum eccentricity for the whole graph.
	Radius	Radius	The minimum eccentricity for the whole graph.
	Center	Center	Number of nodes with an eccentricity value equal to the radius.
	Periphery	Periphery	Number of nodes with an eccentricity value equal to the diameter.
Centrality (9 metrics)	Average in-degree / average out-degree	Degree*	Average of the number of edges converging from / to a node.
	Average neighbor degree	NeighborDegree	The average of the neighbor degree for each distinct node in the graph.
	Average closeness centrality	Closeness*	The average closeness, i.e. centrality of all nodes in the graph.
	Average closeness vitality	Vitality	The average change in closeness for all nodes if successively one node is removed from the graph.
	Average node betweenness centrality	NodeBetweenness	Importance of a node in terms of number of shortest paths passing through this node.
	Average edge betweenness centrality	EdgeBetweenness	Importance of an edge in terms of number of shortest paths passing through this edge.
	Average eigenvector centrality	Eigenvector	Different measures to compute the centrality of a node based on the adjacency matrix of the graph considering the linkage structure of the direct neighborhood of a node and partially a node's own edge structure.
	Average katz centrality	Katz*	
	Average pagerank centrality	Pagerank centrality*	

Table 2. Overview of our applied graph metrics (removed from the final feature set due to multicollinearity, see Chapter 4.2).*

3.3 Prediction Model Training and Assessment

We use prediction models to forecast whether a user session leads to a purchase. We set this target variable to one for all page views in a session if the user conducts a purchase during the session, and to zero otherwise. All predictive variables, i.e. the selected graph metrics and control variables introduced below, are normalized to their standard score to facilitate the interpretation of coefficients for the linear model in terms of their standard deviation from the mean.

We perform out-of-time validation and split our datasets sequentially into training and set; according to the month of the session. Data from September is used as training set whereas data from August is used as test set, resulting in an approximate split of 6:4 between training and test data. Out-of-sample in combination with out-of-time validation is commonly used in benchmarking studies to understand model performance in marketing (Linoff & Berry, 2011, p. 72) or credit scoring (Sobehart et al., 2000), where models are required to be stable over time. This is especially relevant in the e-commerce setting, since we want to test whether our model is able to predict the focal behavior for a different time period than the one in which the model was trained. The out-of-time validation approach is thus stricter in analyzing the performance of the model compared to randomized out-of-sample testing within the same period. We tune the meta-parameters of the prediction models introduced below by means of 5-fold cross validation on the training set. Since our data is highly imbalanced, we additionally applied synthetic minority over-sampling (SMOTE; Chawla et al., 2002), which creates artificial data points based upon the characteristics of a real observation of the minority class and its direct neighborhood to create a balanced dataset.

We select three different classification algorithms, a generalized linear logistic regression model (GLM) and two nonlinear tree-based models, which are random forest (RF) and gradient boosting machine (GB). We motivate the choice of logistic regression by its use in previous work (Byeon, 2013; Kalczynski et al., 2006). RF is chosen due its high performance in several forecasting benchmarks (e.g., Lessmann et al., 2015). We apply GB as a third method, because recent studies have found them to perform superior in similar classification tasks when compared to GLM and RF (Fitzpatrick & Mues, 2016). All models have the advantage that they are interpretable to a degree, which we use to examine the relative predictiveness of alternative graph metrics. The coefficients of logistic regression are interpretable in direction and size and allow significance testing. The RF and GB classifier provides variable importance scores, which also indicate the predictiveness of a variable (Breiman, 2001).

We provide two measures of prediction performance. We evaluate the models build on the respective variable sets using the area-under-the-precision-recall-curves (AUC-PR). The AUC-PR is commonly applied as a single-value metric similar to the area-under-the-ROC-curve (AUC; Fawcett, 2006) for model evaluation in case of imbalanced datasets (Saito & Rehmsmeier, 2015). The PR curve is constructed through pairwise plotting of precision and recall pairs at different classification thresholds, where recall is the proportion of observations predicted to be positive (i.e., purchase) in relation to all positive observations and precision is the rate of predictions that are correct. In general, the higher the value of AUC-PR, the better the model discriminates between the two classes. The second measure we apply is the lift index, which is a popular performance indicator for targeting models (Ling and Li, 1998). Under some assumption, lift is directly connected to the profitability of a targeting model (Martens & Provost, 2011; Piatetsky-Shapiro & Masand, 1999), which further motivates the choice of this performance indicator. The lift is based on a list of customer ordered according to their model-estimate conversion probability. In our case, the lift is defined as the share of hits, i.e. purchasers, in the top segment of $0<\theta<1$ of customers sorted by predicted purchase probability divided by the expected number of buyers in a random sample. More formally, lift L_d is defined as:

$$L_d = \frac{\hat{\pi}_d}{\hat{\pi}} \tag{1}$$

with $\hat{\pi}_d$ denoting the fraction of purchasers among the top-d customers and $\hat{\pi}$ the prior probability of purchase, the lift assesses the degree to which a prediction model improves over a random benchmark.

To be able to assess the performance of our graph-based methodology in comparison to standard approaches, we will use an additional second feature set originating from the standard approach of feature extraction from clickstream (Table 3). Related to Table 1, we will use covariates of different categories such as *Page* and *Time*.

Feature	Description
SessionOverview	Number of visited pages of type 'overview' / 'product'/ 'sale' / 'search' in session.
SessionProduct	
SessionSale	
SessionSearch	
TabVisible	Is the tab currently visible?
Weekday	The weekday the session was started (1 – 7).
DayOfMonth	Day of the month (1 – 31) the session starts.
SessionStartHour	Hour of the session start (morning - midday - evening - night).
TimeOnPage	Time spent on page.
SessionTime	Total time of session.
PageVisitedBefore	Indicator whether the page has been visited before in the session.
Browser	The type of the browser the client uses.
ScreenSize	The screen size resolution of the visitor.
WindowSize	The window resolution of the visitor.
LocationZip	The zip code area of the city the user accesses the website from.
MajorCity	Indicator whether the website access happens from a major city.

Table 3. Overview of traditional features applied as a comparison approach to our graph approach.

4. Empirical Results

Based on the methodology discussed above, we report our empirical results in three steps. Firstly, we will take a detailed look at the correlation among the graph measures applied. Secondly, we analyze the performance of the tested classifiers based on AUC-PR and the lift measure. Finally, we will investigate the different graph measures in order to better understand their impact on the predictive accuracy.

4.1 Dataset Description

We use a two-month period of clickstream data of two large online retailers selling clothing and footwear, respectively. The data was collected from August to September 2015 and contains information such as identifiers (e.g., user id and session id), geographic- and user-based information (e.g., user agent) as well as path-, time- and behavioral-related information with regard to a customer's journey on the respective website.

In the first step, we clean the data by deleting incomplete sessions and dismissing user sessions with less than four page views. Those sessions are referred to as bouncers which have no interest in the website in itself or generally to conduct a purchase. Furthermore, at least four clicks are necessary to complete the purchase process. With regard to potential bot elimination from the dataset we exclude one outlying user sessions with a length of 550 views, which we assume to be the product of automated website access.

The descriptive statistics of the final datasets are shown in Table 4. In total, the first shop contains 58,545 unique users performing a total of 692,975 page views. Of all 80,184 sessions, 4,256 sessions (approx. 5.31%) result in a purchase by a user. The second shop has a lower visitor count of 18,759 users who account for 32,850 sessions and 475,500 page views. Looking at the average value of page views for each visitor, users visiting the second shop on average look at more webpages per session

compared to visitors of the first shop. Still, sessions of website visitors of the first shop result in more purchase conversions than in case of the second shop, where around 0.7 percent less purchases have taken place.

	Shop 1	**Shop 2**
Users	58,545	18,759
Sessions	80,184	32,850
Page views	692,975	475,500
Avg. page views	8.64	14.47
Purchase	4,256 (5.31%)	1,520 (4.63%)

Table 4. Descriptive overview of our final datasets.

4.2 Correlation Analysis of Graph Measures

In the first step, we calculate the correlation matrix of the graph metrics to understand which features embody similar information about the navigational structure of a user's journey on a website. From each set of highly collinear variables, we select only one variable for further analysis to avoid issues of multicollinearity. The corresponding correlation matrices for both datasets are shown in Figure 3. We see that within the three graph metrics categories – structural, distance and centrality – high correlation exists between subsets of the variables. Partly, this is not a surprising result since some measures are either variations of each other (e.g., eigenvector, katz and pagerank centrality) or their calculation is based upon another metric (e.g. eccentricity and the related metrics diameter and radius).

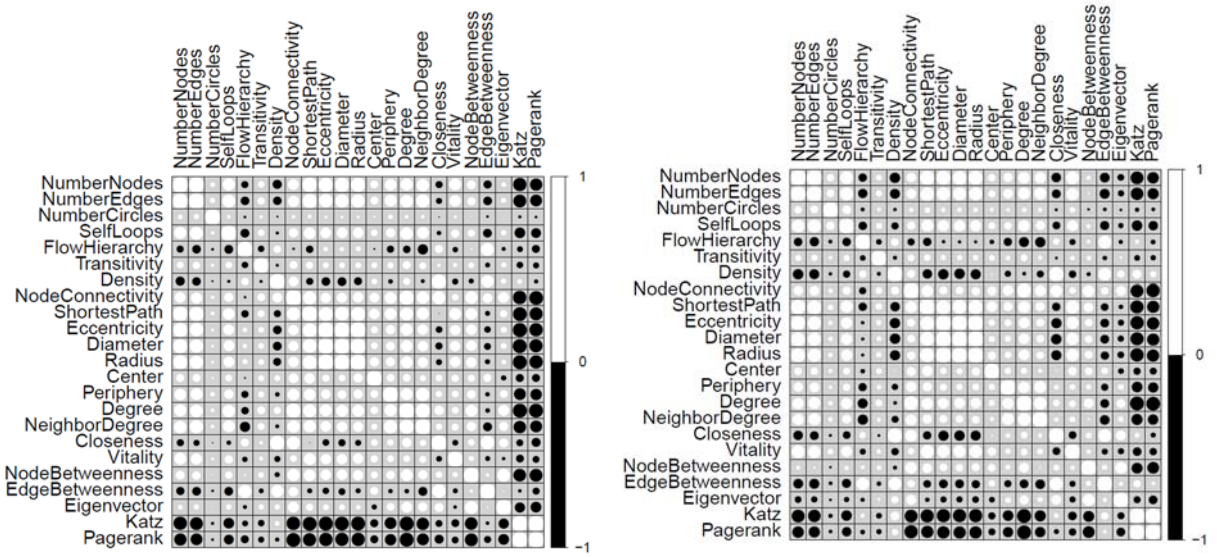

Figure 3. Correlation matrices for shop 1 (left) and shop 2 (right).

For both shops, we observe similar correlation patterns. The measure *NumberNodes* is highly correlated with *NumberEdges*. Additionally, the three metrics *Eccentricity*, *Radius* and *Diameter* contain almost the same informational content. Furthermore, the centrality measures katz and pagerank centrality – themselves highly correlated amongst each other - are negatively correlated with several graph metrics such as the structural components (number of nodes and edges) and the distance-based measures *ShortestPath*, *Eccentricity*, *Diameter* and *Radius*. The correlation of *NumberNodes* and *NumberEdges* can be interpreted in such a way that users tend to often perform as many click events as they visit unique webpages, i.e. webpages are generally visited only once and not several times by a user in a session. The correlation of *Eccentricity*, *Diameter* and *Radius* is unsurprising since they are based on the same basis, i.e. diameter being the maximum and radius being the minimum eccentricity in the graph.

To mitigate the issues of multicollinearity among the graph features, we remove highly correlated features on the basis of their variance inflation factor (VIF; Alin, 2010) calculated as

$$VIF_j = \frac{1}{1-R_j^2}, \quad (2)$$

where R^2 is the coefficient of determination from the regression of the covariate *j* on all other covariates (Stine, 1995). In contrast to the correlation coefficient, the VIF estimates the dependency of one covariate on all other covariates simultaneously, thus avoiding issues of the pairwise comparison. The higher the value of the VIF, the higher the correlation between the covariant *j* and all other variables. In general, covariates exceeding a VIF value between five and ten are seen as being prone to multicollinearity (Katrutsa & Strijov, 2017; Hair et al., 1998). We set our threshold to five and remove covariates exceeding this VIF value from the feature set. The calculation of the VIF is done in a step-wise manner, since the removal of a variable with high correlation affects the remaining variables influence. We recalculate the VIF for all remaining variables after removing the covariant with the highest VIF value from the preceding evaluation round. For both shops VIF results were almost consistent leading to the elimination of the covariates *Katz* (VIF = 2536.10 for shop 1 / 1201.07 for shop 2), *Diameter* (VIF = 176.74 / 758.24), *NumberNodes* (VIF = 166.33 / 200.10), *NodeConnectivity* (VIF = 82.06 / 45.88), *Closeness* (VIF = 30.37 / 22.65), *Pagerank* (VIF = 25.31 / 15.43), *ShortestPath* (VIF = 14.97 / 10.42), *Degree* (VIF = 10.59 / 5.08) and *NumberEdges* (VIF = 7.02 / 5.08) from the feature set. Additionally, in case of shop 1 the covariant *Eccentricity* (VIF = 993.34) and in case of shop 2 the *Radius* (VIF = 264.23) exceed the VIF threshold. Since these two metrics show high VIF values from the very beginning which drop significantly once either of the two is removed, we remove the covariant with the higher overall VIF, *Eccentricity,* (VIF = 993.34 for shop 1 / 257.25 for shop 2) for both shops and keep the covariant *Radius* in both datasets to increase consistency and facilitate the analysis.

This results in a final feature set consisting of 13 graph metrics, which we use for further analysis.

4.3 Predictive Performance

Using the subset of the 13 remaining graph features, we compare their predictive performance against the traditional feature set based on the GLM, RF and GB algorithms introduced above.

Looking at AUC-PR (Table 5), we observe that the graph-based approach outperforms the traditional set of variables in all six instances independent of the underlying model. We further observe that the RF performs worse compared to GB and the linear GLM for both shops. Furthermore, apart from the RF model, both models achieve higher AUC-PR values in case of shop 1. All models outperform the expected performance of a random model equal to the purchase rate of 5.3% and 4.6% for shop 1 and 2, respectively.

Model	**GLM**		**RF**		**GB**	
Covariates	**Graph**	**Traditional**	**Graph**	**Traditional**	**Graph**	**Traditional**
Shop 1	**0.372**	0.271	**0.287**	0.262	**0.372**	0.262
Shop 2	**0.311**	0.243	**0.300**	0.247	**0.317**	0.288

Table 5. AUC-PR values for shop 1 and shop 2 for the applied models.

For the lift measure, we observe that all three models trained on the applied graph metrics constitute for a clear improvement compared to random targeting. Figure 4 visualizes model lift in a gain chart for the three models on each dataset. Intuitively, the gain chart provides information about the number of purchasers if n% of users are targeted by the model, for example with a marketing incentive. Along the x-axis all views are plotted ordered by their predicted probability to purchase starting with those views having the highest probability. The y-axis represents the cumulative number of purchases among those page views. The upper grey bound shows the outcome of a perfect model which classifies all views according to their correct outcome, while a random model would result in a 45-degree diagonal. The steeper the curve is for a model, the better the model.

In case of the first shop, for the first around 30 percent of samples tested, both the GLM and the GB model perform almost equally, i.e. their performance in terms of classifying views with a high predicted purchase probability. In the beginning, RF performs slightly worse until around 30 percent of the samples are tested where the model exhibits a similar performance compared to the other two applied models but is soon visibly outperformed for larger samples. For the second shop, all three models are even more homogenous in terms of their predictive performance until a threshold of around 50 percent of samples is reached. Exceeding this threshold we observe again a similar performance of GLM and GB, while the RF model falls behind.

In general, the GLM model performs comparable in terms of lift compared to GB. This is surprising considering the general performance of the models and the ability of GB to model non-linear relations between the predictors. Given that all graph metrics are different measures to describe the same underlying graph structure, the good performance of the logistic regression model might be an indication that there are no significant non-linear dependencies between the graph metrics in predicting purchase behavior

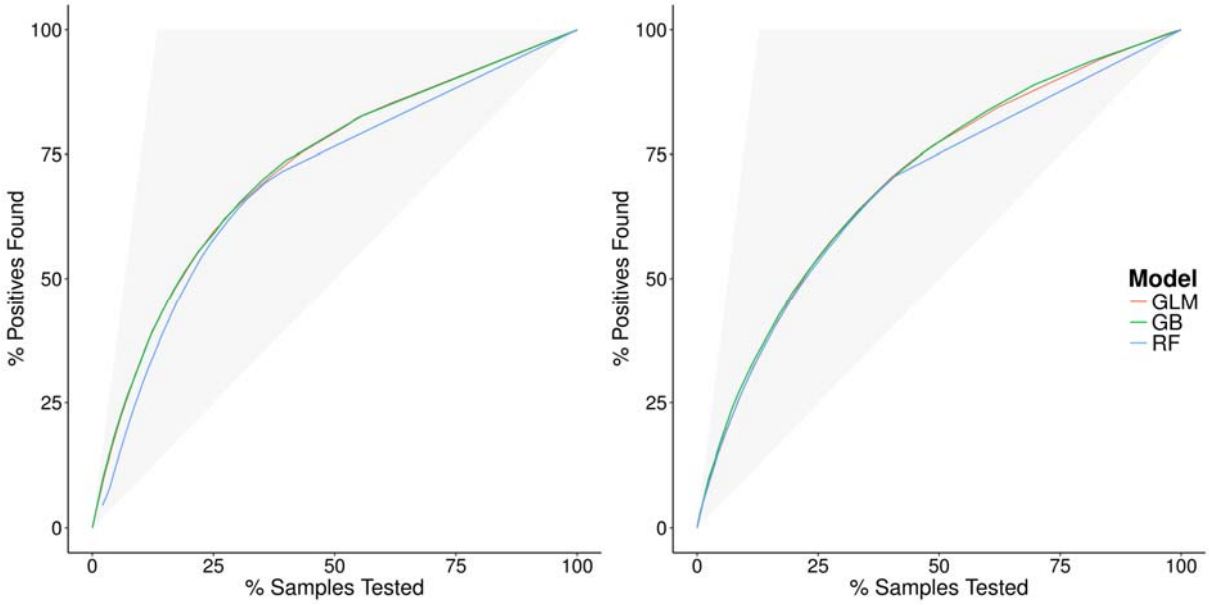

Figure 4. Lift chart for shop 1 (left) and shop 2 (right).

4.4 Variable Importance

In order to shed light on the direction of effect and performance of each graph metric, we analyze the model-wise importance of each graph measure. For the GLM model, we report the raw coefficients and odds ratios for each dataset (Table 6). Due to the large data size, we observe that almost all coefficients are highly significant even at the 0.01% level. The coefficient of the variable *NumberCircles* is least significant at the 0.1% level. We thus focus on the analysis of effect size.

Since all variables are standardized, we analyze their effect size expressed in terms of the impact a change by one standard deviation (SD) has on the odds ratio (Table 6). With the odds ratio defined by the ratio of probabilities $\mathrm{P}(Purchase)/\mathrm{P}(No\ purchase)$, an exponentiated effect above 1 indicates a larger purchase probability. For both shops, *Radius* and *SelfLoops* have a strong positive effect on purchase probability. Specifically, an increase in *Radius* by one standard deviation, associated with less compact graphs, indicates an increase in purchase odds by 135% (shop 1) or 51% (shop 2), while an increase in *SelfLoops* by one SD leads to an increase in purchase odds by 40% (shop 1) or 51% (shop 2). A slightly smaller effect exists for *EdgeBetweenness* where a one SD increase, due to less connections between nodes, is associated with a 12% (shop 1) or 28% (shop 2) increase. In contrast, we observe the largest negative impact on purchase odds for a decrease in *Density*, where a decrease

by one SD, observed for sparser graphs, increases the odds of a purchase by 23% (1/0.81 = 1.23) (shop 1) or 39% (shop 2). *FlowHierarchy* is estimated to have a negative effect of similar size. While there are some differences in effect size, we observe no difference in direction for the above variables, which have the strongest impact. In sum, the observed pattern suggests that linear click-paths related to search behavior may be more indicative of users with purchase intention.

The variables *NumberCircles, Eigenvector NeighborDegree,* and *Center* show coefficients in different directions between shop 1 and 2, indicating that the underlying relationship may be shop dependent to a larger degree.

	GLM Model for Shop 1			**GLM Model for Shop 2**		
Variable	**Coefficient**	**Std. Error**	**Odds Ratio**	**Coefficient**	**Std. Error**	**Odds Ratio**
Intercept	-0.29 ***	0.004	0.75	-0.29 ***	0.005	0.75
NumberCircles	-0.04 **	0.014	0.96	0.09 ***	0.013	1.09
Density	-0.21 ***	0.006	0.81	-0.33 ***	0.009	0.72
Vitality	-0.12 ***	0.006	0.89	-0.06 ***	0.009	0.94
NodeBetweenness	0.05 ***	0.005	1.05	0.03 ***	0.006	1.03
EdgeBetweenness	0.11 ***	0.008	1.12	0.25 ***	0.010	1.28
Eigenvector	-0.06 ***	0.005	0.94	0.02 **	0.006	1.02
Radius	0.86 ***	0.007	2.35	0.41 ***	0.008	1.51
SelfLoops	0.34 ***	0.005	1.40	0.51 ***	0.007	1.66
FlowHierarchy	-0.20 ***	0.006	0.82	-0.25 ***	0.007	0.78
NeighborDegree	-0.15 ***	0.007	0.86	0.05 ***	0.006	1.06
Center	-0.10 ***	0.005	0.91	0.03 ***	0.006	1.03
Periphery	0.14 ***	0.005	1.15	0.05 ***	0.005	1.05
Transitivity	0.08 ***	0.004	1.09	0.11 ***	0.005	1.11
Significance levels: 0.0001 '***' 0.001 '**' 0.01 '*'						

Table 6. Estimated coefficients for the GLM model.

For the gradient boosted trees, we calculate the variable bag importance for both datasets based on the weighted increase in node purity for the splits on each variable averaged over all trees (Hastie et al., 2001). In other words, the variable importance captures the relative contribution to improve classification for each variable in the model. The variable importance scores reported in Figure 5 are scaled to sum up to 100 and are ordered according to their average importance for both shops.

The importance ranking for the non-linear GB model shows different patterns compared to the logit coefficient analysis in so far as *Density* and *FlowHierarchy* are only marginally relevant for purchase prediction while *Vitality*, *SelfLoops*, *NumberCircles* and *Radius* constitute the most important variables. Since high values of *Vitality* refer to the existence of important connections in the graph structure, the importance of the variable could be explained through being able to detect specific user behavior, i.e. signifying either goal-oriented, non-recursive browsing behavior or the existence of bridging elements in the user website journey such as overview or search result pages.

However, we observe that feature importance is centered around *Vitality* with a sharp decrease towards the second most important variable. Here we observe some deviation in variable importance between shops. In case of shop 1 the variables *Radius*, *Periphery* and *SelfLoops* are the next most important variables, whereas in case of shop 2 this is true for the variables *SelfLoops*, *NumberCircles* and *Center*. While the distance-based measures *Radius*, *Periphery* and *Center* refer to global characteristics of the clickstream graph, *SelfLoops* and *NumberCircles* are related to direct

neighborhoods of single nodes. These two feature pairs could flag different browsing behaviors present in both shops.

All other variables only constitute for a small percentage in terms of variable importance and seem to be negligible for the distinction of purchasers and non-purchasers in case of the two datasets applied and the GB model.

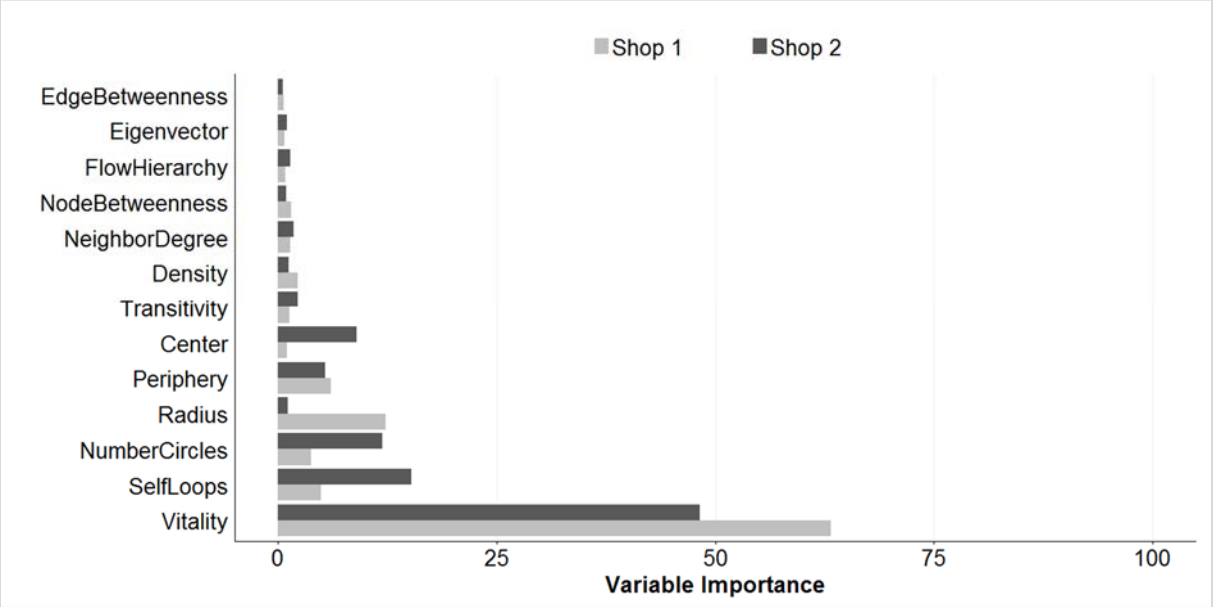

Figure 5. Variable importance for GB model for shop 1 and shop 2.

We use Partial Dependence Plots (PDP) for a deeper analysis of non-linear effects of each variable within the GB model (Hastie et al., 2001). PDP are a graphical tool to examine the marginal effect of each variable on the model prediction accounting for the (average) effects of all other variables. Figure 6 and Figure 7 show the PDP for shop 1 and shop 2, respectively. For both shops, we partly observe distinctive patterns, not completely in line with the general between-shops robustness of effects observed for GLM. Naturally, the PDP for both shops show the most distinctive patterns for the variables with the highest important scores (see Figure 5).

According to the PDP, an increasing value of *Vitality*, being the most important variable for both shops and which we interpret as a rising number of central pages in the user journey, shows similar behavior for both shops. Initially values below zero are linked to a high purchase probability but drop significantly once a value of zero is reached. However, after *Vitality* reaches a certain threshold the purchase probability increases again for both shops. Since this measure represents the change of distances for all present nodes in the graph, this metric might both be able to capture non-recursive type of browsing behavior and users with a high number of page views, both possibly being an indication of shopping behavior leading to a purchase.

The PDP of *SelfLoops*, which captures re-occurring webpage visits and constitutes the second most important variable, reveals a similar link. For both shops the purchase probability increases with the number of times a user revisits the same page. However, whereas in case of shop 1 the purchase probability continuously increases with a larger number of loops in the browser session this is not the case for shop 2. After reaching a value of around four the purchase probability decreases and remains stable from then on. Furthermore, an increasing value of *NumberCircles*, representing the existence of circles in the browsing structure of users, leads to an increasing purchase probability for both shops.

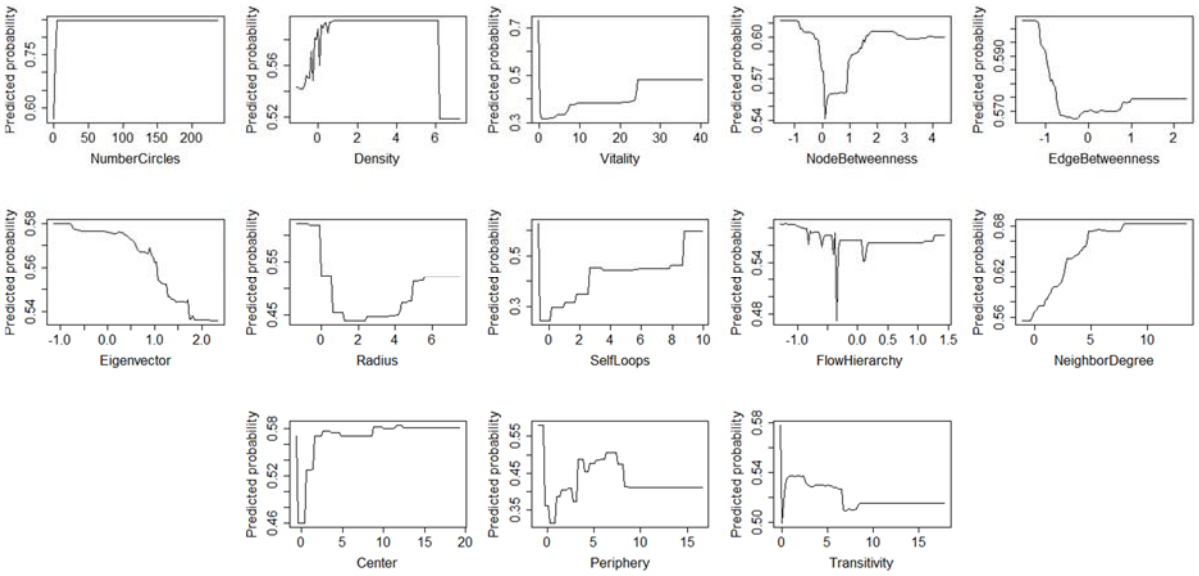

Figure 6. Partial dependence plots for shop 1.

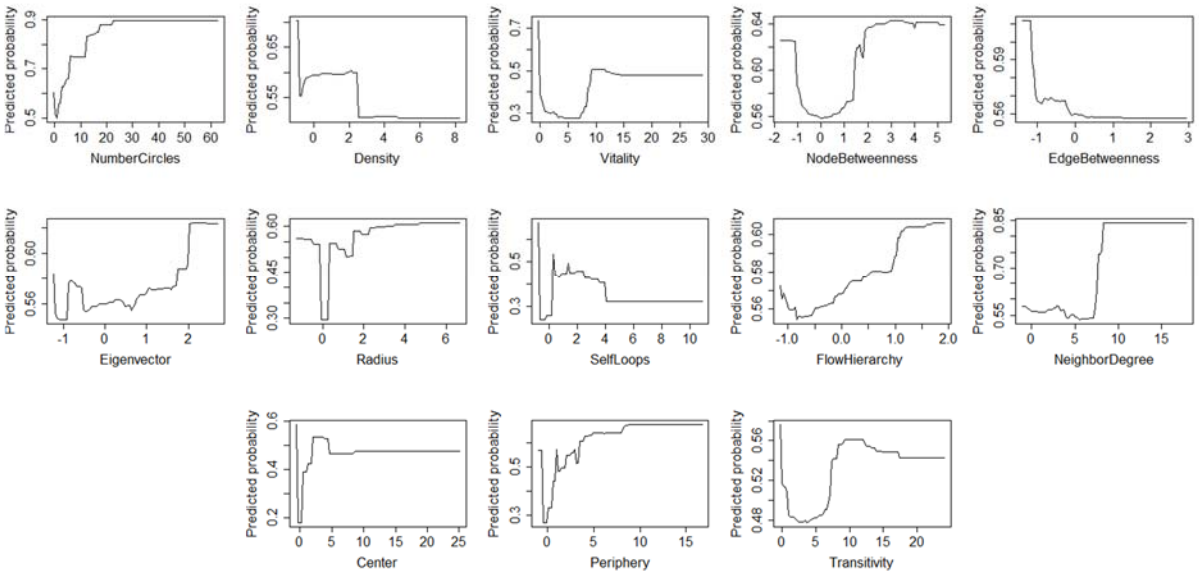

Figure 7. Partial dependence plots for shop 2.

In case of shop 1, the two distance-based measures *Radius* and *Periphery* have been shown to be important for the prediction task. In general, the PDPs of both metrics resemble a U-shaped curve where both low and high values are associated with a higher purchase probability. For both metrics a drop occurs at a value of around zero where purchase probability resembles the lowest value.

In case of shop 2, the variable importance scores of *Center* and *Periphery* have been shown to be relevant for predicting purchasers as captured by the feature important scores. Despite a sharp drop at a value of zero, the PDP for *Center* illustrates the general relationship that the higher the value of this metric the more likely it is that a purchase occurs, signaling extensive browsing behavior within a direct neighborhood of a webpage. For *Transitivity* again a U-shaped curve is observable where the valley represents those users who generally do not conduct a purchase. Furthermore, rather low and values exceeding a value of five might signal two different shopping behaviors leading to a purchase which are goal-oriented and browsing-related shopping behavior.

Altogether these might be indicators that graph metrics are able to detect different shopping behaviors leading to a purchase which are either a goal-oriented or a browsing-related shopping experience. However, given the predictive value of the graph metrics, in-depth analysis beyond the scope of this paper will be necessary to identify the specific user intentions associated with a graph structure and could focus on experimental investigation of the link between stated user intention and each metric and establishing the robustness of the observed dependencies structures to different shops and product categories.

5. Conclusion

Using real-life clickstream datasets of two different shops we observe for both the linear GLM model and the non-linear Random Forest model that distance- and centrality-based graph metrics are effective in predicting purchase behavior of users. We derived user-centered, session-based graphs from clickstream data, where each graph is developed incrementally, i.e. each new page view of the user develops the graph further. Each of the 23 tested graph metrics are calculated for each intermediate state of a graph. We report and control for multicollinearity between the graph metrics by pre-processing using variable inflation factors and train three selected high-performing algorithms on the resulting dataset. Independent of the employed model, the proposed variables result in a substantial increase in the area-under-the-precision-recall-curve and model lift in predictive power compared to random targeting and a set of standard aggregation features derived from clickstream.

Looking at the importance of each graph metric, we observe clear differences in the relevance of variables between the linear and non-linear models. We suggest that closeness vitality in particular followed by radius and the number of self-loops and circles should be considered promising candidates in future applications.

We also identify some promising areas for future research. An alternative approach to calculate graph metrics could include different graph construction methods such as using bi-partite graphs, where two different types of nodes are included, to represent the structure of a user session in more detail. Additionally, constructing weighted graphs by rating frequently taken paths as more important or accounting for the time spent on specific pages could improve the representation of the users' journey on a website and consequently increase the accuracy when predicting the outcome of a session.

References

Alin, A. (2010). Multicollinearity. Wiley Interdisciplinary Reviews: *Computational Statistics*, 2(3), 370-374.

Anitha, A. (2010). A new web usage mining approach for next page access prediction. *International Journal of Computer Applications*, 8(11), 7-10. doi:10.5120/1252-1700

Antonellis, P., Makris, C., & Tsirakis, N. (2009). Algorithms for clustering clickstream data. *Information Processing Letters*, 109(8), 381-385. doi:10.1016/j.ipl.2008.12.011

Banerjee, A., & Ghosh, J. (2001). Clickstream clustering using weighted longest common subsequences. *Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining* (pp. 33-40).

Berka, P., & Labský, M. (2007). Predicting page occurrence in a click-stream data: statistical and rule-based approach. *Advances in Data Mining. Theoretical Aspects and Applications* (pp. 135-147). Springer Berlin Heidelberg.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.

Byeon, H. (2013). Evaluating the online buying behavior using network analysis. *International Journal of Advancements in Computing Technology*, 5(12).

Chan, T., Joseph, I., Macasaet, C., Kang, D., Hardy, R. M., Ruiz, C., Porras, R., Baron, B., Qazi, K., Hannon, P. & Honda, T. (2014). Predictive models for determining if and when to display online lead forms. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (pp. 2882-2889). AAAI Press.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: synthetic minority over-sampling technique*. Journal of artificial intelligence research, 16, 321-357.

Colladon, A. F., & Remondi, E. (2017). Using social network analysis to prevent money laundering. *Expert Systems with Applications*, 67, 49-58. doi: 10.1016/j.eswa.2016.09.029

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874. doi:10.1016/j.patrec.2005.10.010

Fitzpatrick, T., & Mues, C. (2016). An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market. *European Journal of Operational Research*, 249(2), 427-439. doi:10.1016/j.ejor.2015.09.014

Girija, P., & Kavitha, V. (2013). An Approach for predicting user's web access pattern. *International Journal of Computer Science and Management Research*, 2(5), 2585-2589.

Gündüz, Ş., & Özsu, M. T. (2003). A web page prediction model based on click-stream tree representation of user behavior. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 535-540). ACM.

Hagberg, A. A., Schult, D. A. & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX", *Proceedings of the 7th Python in Science Conference* (pp. 11–15*)*, Gäel Varoquaux, G., Vaught, T. & Millman, J. (Eds), Pasadena, CA USA.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). *Multivariate data analysis,* 5(3), Upper Saddle River, NJ: Prentice hall, 207-219.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning* (2nd ed.). New York: Springer series in statistics.

He, Y. L., Liu, J. N., Hu, Y. X., & Wang, X. Z. (2015). OWA operator based link prediction ensemble for social network. *Expert Systems with Applications*, 42(1), 21-50. doi: 10.1016/j.eswa.2014.07.018

Hong, T., & Kim, E. (2012). Segmenting customers in online stores based on factors that affect the customer's intention to purchase. *Expert Systems with Applications*, 39(2), 2127-2131. doi: 10.1016/j.eswa.2011.07.114

Iwanaga, J., Nishimura, N., Sukegawa, N., & Takano, Y. (2016). Estimating product-choice probabilities from recency and frequency of page views. *Knowledge-Based Systems*, 99, 157-167. doi:10.1016/j.knosys.2016.02.006

Jiang, Q., Tan, C. H., & Wie, K. K. (2012). Cross-website navigation behavior and purchase commitment: A pluralistic field research. *Proceedings of the 16th Pacific Asia Conference on Information Systems (PACIS)*.

Kalczynski, P. J., Senecal, S., & Nantel, J. (2006). Predicting on-line task completion with clickstream complexity measures: A graph-based approach. *International Journal of Electronic Commerce*, 10(3), 121-141. doi:10.2753/jec1086-4415100305

Katrutsa, A., & Strijov, V. (2017). Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Systems with Applications*, 76, 1-11.

Kitts, B., Hetherington-Young, K., & Vrieze, M. (2002). Large-scale mining, discovery and visualization of WWW user clickpaths. *International Journal of Image and Graphics*, 02(01), 21-48. doi:10.1142/s0219467802000536

Lee, H., Choi, S. Y., & Kang, Y. S. (2009). Formation of e-satisfaction and repurchase intention: Moderating roles of computer self-efficacy and computer anxiety. *Expert Systems with Applications*, 36(4), 7848-7859. doi:10.1016/j.eswa.2008.11.005

Lee, M., Ferguson, M. E., Garrow, L. A., & Post, D. (2010). The impact of leisure travelers' online search and purchase behaviors on promotion effectiveness. Working Paper, Georgia Institute of Technology.

Lessmann, S., Baesens, B., Seow, H., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124-136. doi:10.1016/j.ejor.2015.05.030

Ling, C. X., & Li, C. (1998). Data mining for direct marketing: Problems and solutions. In R. Agrawal, P. E. Stolorz, G. Piatetsky-Shapiro (Eds.). *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining* (pp. 73–79). AAAI Press, Menlo Park.

Linoff , G. S. & Berry, M. J. A. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (3rd ed.). Wiley Publishing.

Lu, L., Dunham, M., & Meng, Y. (2005). Mining significant usage patterns from clickstream data. *Advances in Web Mining and Web Usage Analysis* (pp. 1-17). Springer Berlin Heidelberg.

Martens, D., & Provost, F. (2011). Pseudo-social network targeting from consumer transaction data. Faculty of Applied Economics, University of Antwerp.

Moe, W. W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology*, 13(1-2), 29-39. doi:10.1207/s15327663jcp13-1&2_03

Moe, W. W., & Fader, P. S. (2004). Capturing evolving visit behavior in clickstream data. *Journal of Interactive Marketing*, 18(1), 5-19. doi:10.1002/dir.10074

Moe, W. W., Chipman, H., George, E. I., & McCulloch, R. E. (2002). A Bayesian treed model of online purchasing behavior using in-store navigational clickstream. Working paper, University of Texas at Austin, Austin.

Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data. *Marketing Science*, 23(4), 579-595. doi: 10.1287/mksc.1040.0073

Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., & Vanthienen, J. (2017). Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. To appear in: *Expert Systems with Applications*. doi: 10.1016/j.eswa.2017.05.028

Padmanabhan, B., Zheng, Z., & Kimbrough, S. O. (2006). An empirical analysis of the value of complete information for eCRM models. *MIS Quarterly*, 247-267.

Pai, D., Sharang, A., Yadagiri, M. M., & Agrawal, S. (2014). Modelling visit similarity using click-stream data: A supervised approach. *Web Information Systems Engineering – WISE* (pp. 135-145). Springer International Publishing.

Panagiotelis, A., Smith, M. S., & Danaher, P. J. (2014). From Amazon to Apple: Modeling online retail sales, purchase incidence, and visit behavior. *Journal of Business & Economic Statistics*, 32(1), 14-29. doi: 10.1080/07350015.2013.835729

Park, C. H., & Park, Y. (2016). Investigating purchase conversion by uncovering online visit patterns. *Marketing Science*, 35(6), 894-914. doi:10.1287/mksc.2016.0990

Park, S., Suresh, N. C., & Jeong, B. K. (2008). Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm. Data & Knowledge Engineering, 65(3), 512-543. doi: https://doi.org/10.1016/j.datak.2008.01.002

Piatetsky-Shapiro, G., & Masan, B. (1999). Estimating campaign benefits and modeling lift. *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*. San Diego, CA, USA.

Pitman, A., & Zanker, M. (2010). Insights from applying sequential pattern mining to e-commerce click stream data." In: *IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 967-975). IEEE.

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One*, 10(3), e011843.

Sarwar, S. M., Hasan, M., & Ignatov, D. I. (2015). Two-stage cascaded classifier for purchase prediction." arXiv preprint arXiv:1508.03856.

Sato, S., & Asahi, Y. (2012). A daily-level purchasing model at an e-commerce site. *International Journal of Electrical and Computer Engineering*, 2(6). doi:10.11591/ijece.v2i6.1816

Senecal, S., Kalczynski, P. J., & Fredette, M. (2014). Dynamic identification of anonymous consumers visit goals using clickstream. *International Journal of Electronic Business*, 11(3), 220. doi:10.1504/ijeb.2014.063036

Shams, B., & Haratizadeh, S. (2017). Graph-based collaborative ranking. *Expert Systems with Applications*, 67, 59-70. doi:10.1016/j.eswa.2016.09.013

Sismeiro, C., & Bucklin, R. E. (2004). Modeling purchase behavior at an e-commerce web site: A task-completion approach. *Journal of Marketing Research*, 41(3), 306-323. doi:10.1509/jmkr.41.3.306.35985

Stange, M., & Funk, B. (2015). How much tracking is necessary? – The learning curve in bayesian user journey analysis. *Proceedings of the 23rd European Conference on Information Systems*. Muenster, Germany.

Statista (2016a). Global retail e-commerce sales 2014-2020 | Statistic. Retrieved May 22, 2017, from http://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/

Statista (2016b). Global e-retail growth rate 2020 | Statistic. Retrieved May 22, 2017, from http://www.statista.com/statistics/288487/forecast-of-global-b2c-e-commerce-growt/

Stine, R. A. (1995). Graphical interpretation of variance inflation factors. *The American Statistician*, 49(1), 53-57.

Suh, E., Lim, S., Hwang, H., & Kim, S. (2004). A prediction model for the purchase probability of anonymous customers to support real time web marketing: a case study. *Expert Systems with Applications*, 27(2), 245-255. doi:10.1016/j.eswa.2004.01.008

Van den Poel, D., & Buckinx, W. (2005). Predicting online-purchasing behaviour. *European Journal of Operational Research*, 166(2), 557-575. doi:10.1016/j.ejor.2004.04.022

Van der Meer, D., Dutta, K., Datta, A., Ramamritham, K., & Navanthe, S. B. (2000). Enabling scalable online personalization on the web. *Proceedings of the 2nd ACM Conference on Electronic Commerce* (pp. 185-196). ACM.

Vroomen, B., Donkers, B., Verhoef, P. C., & Franses, P. H. (2005). Selecting profitable customers for complex services on the Internet. *Journal of Service Research*, 8(1), 37-47. doi:10.1177/1094670505276681

Wu, F., Chiu, I. H., & Lin, J. R. (2005). Prediction of the intention of purchase of the user surfing on the web using hidden markov model. In: *Proceedings of ICSSSM'05 International Conference on Services Systems and Services Management.* IEEE, pp. 387-390.

Zhang, Y., Bradlow, E. T., & Small, D. S. (2015). Predicting customer value using clumpiness: From RFM to RFMC. *Marketing Science*, 34(2), 195-208. doi:10.1287/mksc.2014.0873

Zhao, Y., Yao, L., & Zhang, Y. (2016). Purchase prediction using Tmall-specific features. *Concurrency and Computation: Practice and Experience*, 28(14), 3879-3894. doi:10.1002/cpe.3720

Zheng, Z., Padmanabhan, B., & Kimbrough, S. O. (2003). On the existence and significance of data preprocessing biases in web-usage mining. *INFORMS Journal on Computing*, 15(2), 148-170. doi:10.1287/ijoc.15.2.148.14449

Appendix

Shop 1							
	Min.	**25% Quant.**	**Median**	**Mean**	**75% Quant.**	**Max.**	**Std. Dev.**
Purchase	0.00			0.14		1.00	
NumberCircles	0.00	0.00	1.00	3.36	3.00	38150.00	132.95
Density	0.00	0.13	0.25	0.30	0.42	2.00	0.26
Vitality	0.00	1.00	7.50	54.51	36.00	25818.68	289.35
NodeBetweenness	0.00	0.00	0.15	0.12	0.17	0.50	0.09
EdgeBetweenness	0.00	0.13	0.20	0.21	0.28	0.50	0.14
Eigenvector	0.00	0.00	0.21	0.19	0.33	0.58	0.18
Radius	0.00	1.00	2.00	1.79	2.00	24.00	1.38
SelfLoops	0.00	0.00	0.00	0.67	1.00	13.00	1.02
FlowHierarchy	0.00	0.13	0.44	0.49	1.00	1.00	0.38
NeighborDegree	0.00	0.33	1.00	1.11	1.63	23.91	1.09
Center	1.00	1.00	1.00	1.58	2.00	20.00	0.85
Periphery	1.00	2.00	2.00	2.49	3.00	30.00	1.49
Transitivity	0.00	0.00	0.00	0.01	0.00	1.00	0.05

A 1. Summary statistics of graph metrics for shop 1.

Shop 2							
	Min.	**25% Quant.**	**Median**	**Mean**	**75% Quant.**	**Max.**	**Std. Dev.**
Purchase	0.00			0.13		1.00	
NumberCircles	0.00	1.00	2.00	7.49	7.00	14298.00	117.34
Density	0.00	0.09	0.18	0.25	0.33	2.00	0.23
Vitality	0.00	4.00	24.29	225.33	125.17	29824.62	897.00
NodeBetweenness	0.00	0.09	0.14	0.12	0.17	0.50	0.08
EdgeBetweenness	0.00	0.09	0.15	0.18	0.22	0.50	0.12
Eigenvector	0.00	0.00	0.18	0.19	0.32	0.58	0.16
Radius	0.00	1.00	2.00	2.76	4.00	28.00	2.32
SelfLoops	0.00	0.00	1.00	1.09	1.00	37.00	1.64
FlowHierarchy	0.00	0.12	0.33	0.42	0.67	1.00	0.35
NeighborDegree	0.00	0.80	1.39	1.79	2.23	48.61	2.03
Center	1.00	1.00	1.00	1.62	2.00	31.00	1.03
Periphery	1.00	2.00	2.00	2.89	3.00	39.00	1.98
Transitivity	0.00	0.00	0.00	0.01	0.00	1.00	0.04

A 2. Summary statistics of graph metrics for shop 2.