Visual exploration of rating datasets and user groups

Fabian Colque Zegarra, Juan Carbajal Ipenza, Behrooz Omidvar-Tehrani, Viviane Moreira, Sihem Amer-Yahia, João L.D. Comba

Visual Exploration of Rating Datasets and User Groups Authors' Copy

Fabian Colque Zegarra[a], Juan C. Carbajal[a], Behrooz Omidvar-Tehrani[b], Viviane Moreira[a], Sihem Amer-Yahia[c,d], João L. D. Comba[a]

[a] *UFRGS (Brazil)*
[b] *NAVER LABS Europe (France)*
[c] *University of Grenoble Alpes (France)*
[d] *CNRS (France)*

Abstract

The increasing availability of rating datasets (*i.e.,* datasets containing user evaluations on items such as products and services) constitutes a new opportunity in various applications ranging from behavioral analytics to recommendations. In this paper, we describe the design of VUGA, a visual enabler for the exploration of rating data and user groups. VUGA helps analysts, be they novice analysts or domain experts, acquire an understanding of their data through a seamless integration between exploring users and exploring their collective behavior via group analysis. VUGA is data-driven and does not require analysts to know the value distributions in their data. While automated systems can identify and suggest potentially interesting groups, they can do that for well-specified needs (*e.g.,* through SQL QUERIES or constrained mining). VUGA helps analysts *filter and refine their exploration as they discover what lies in the data.* VUGA enables analysts to *easily acquire statistics about their data, form groups, and find similar and dissimilar groups.* While most visual analytics systems are data-dependent, VUGA relies on a data model that captures user data in such a way that a variety of group formation and exploration approaches can be used.

Email addresses: feczegarra@inf.ufrgs.br (Fabian Colque Zegarra), juan.carbajalipenza@inf.ufrgs.br (Juan C. Carbajal), behrooz.omidvar-tehrani@naverlabs.com (Behrooz Omidvar-Tehrani), viviane@inf.ufrgs.br (Viviane Moreira), sihem.amer-yahia@univ-grenoble-alpes.fr (Sihem Amer-Yahia), comba@inf.ufrgs.br (João L. D. Comba)

Preprint submitted to Special Issue on "Data Exploration in the Web 3.0 Age" December 1, 2020

We describe the architecture of VUGA and illustrate its use via tasks and a user study. We conclude with a discussion on future work enabled by VUGA.

Keywords: User Data Exploration, User Group Exploration, Visual Analytics, User Data, Dimensionality Reduction

1. Introduction

Scientists and non-scientists increasingly rely on user data to achieve a variety of tasks with the target goal of finding people of interest or analyzing collective behavior. Many examples of user data can be found on the social web and more particularly in online rating systems. In general terms, rating datasets can be described by a combination of *demographics* (*e.g.,* age, gender, occupation) and *events* (*e.g.,* movie or book ratings). Given their high volume, understanding what lies behind those datasets is a daunting task. More specifically, identifying group behavior in those datasets relies on the ability to explore the space of users, aggregate their demographics and interests, and compare groups. While automated systems can identify and suggest potentially interesting groups, the need for an interactive process that provides filtering capabilities requires a visual interface. The combination of a visual interface and statistics with filtering capabilities is referred to as a Visual Analytics (VA). In this paper, we describe VUGA, a VA system for exploring users and forming and exploring groups in datasets that have demographics and ratings.

A visual user group analytics system integrates *user data exploration* and *group exploration* via a visual interface. User data exploration is the process through which analysts acquire an understanding of users and their statistics [1]. This process relies on the ability to visually filter users based on common demographics or interests. This process results in a seed group that is built by the analyst for further exploration. Group exploration takes as input a seed group and returns similar and dissimilar groups [2]. User and group explorations both serve scenarios where analysts express partial needs and require to build on the knowledge they acquire as they see more data. Providing an interface

enables visual inspection of the data and lets the analyst intervene to apply filters, handpick users in groups, form new groups, and request related groups. VUGA caters to analysts with varying levels of expertise. *Novice analysts* are generally interested in completing daily tasks such as finding a movie or starting a book club. For that, they need to find people like them and alternate between a user-centric view and a group-centric view of the data. They also need to explore individual and collective interests to reach a decision. *Domain experts,* on the other hand, tend to look for validating assumptions on their data, the so-called "Confirmatory Analysis" [3]. For instance, they want to verify if middle-aged people prefer Drama movies over Comedies. For that, they need to obtain a holistic view of statistics and data distributions associated with demographics and movie genres. VUGA offers *an integrated system* within which *both users and groups* can be explored.

While the status quo of analyzing user data and user groups is to get statistics, form groups, and explore them with separate tools [4], VUGA combines the power of a visual interface with an interactive exploration of user data and groups. This has both transfer and time overheads [5]. Some existing work alleviates that by providing partial handshaking between different components. For instance, visual interfaces are integrated with user exploration [6] and group exploration [7, 8, 9, 10]. There are also many exploration methods without a visual interface [11, 12, 13, 14]. VUGA is a fully-integrated visual exploration system for rating data.

VUGA relies on a data model that captures a variety of user data and additional information that allows to compose and iterate between the user and group explorations. The model represents a user as a single vector that gathers demographics and interests. A set of users is then a set of vectors to which a 2D projection is applied for visualization. The projection is based on computing vector similarity between user vectors and visualizing similar users closer to each other. The 2D projection is done using t-distributed Stochastic Neighbor Embedding technique (t-SNE) [15], which is widely applied in machine learning, visual analytics, and data mining. This projection enables analysts to form a

seed group which is fed to group exploration to find similar/dissimilar groups. This interleave between modeling users and groups enables seamless handshaking between user and group explorations through a visual interface.

Contributions. VUGA makes the following contributions.

1. The ability to represent, ingest and visualize a variety of user and rating data in a generic fashion.

2. The ability to visually filter and form user groups on-the-fly along multiple dimensions. Additionally, the visualization is enriched with a coordinated view of various statistics associated with groups.

3. The ability to use a group exploration method that takes a seed group and find similar/dissimilar groups through seamless integration between a user-centric and a group-centric view of user data.

4. Support for different use cases that cater to analysts with different data expertise. Two representative use cases and a user study that validate the need for an integrated system for user and group explorations.

Outline. Section 2 describes the components of VUGA and positions our contributions with respect to related work. The design considerations of a visual exploration system for user data are outlined in Section 3. The overall architecture of VUGA is outlined in Section 4. A detailed description of VUGA in action is described in Section 5. We present a user study in Section 6. Section 7 concludes with a summary and a discussion of future directions.

2. Related Work

Exploring rating data calls for an understanding of data about users and the ability to handpick users of interest (referred to as group formation). A group is a set of individuals with common demographics and events. Forming such groups enables analysts to understand the collective behavior of individuals in groups.

Often there exist millions of groups in user data which put a burden on analysts to pick groups of interest manually. Group exploration methods help analysts to navigate in the plethora of groups in an effective way. Visualization helps analysts make sense of explored data and groups. VUGA is a mixed-initiative framework [16] which incorporates rating data exploration, group exploration, and visualization, in a fully connected fashion. We structure the related work by the features and analytical tasks that VUGA provides. First, we review visual analytics approaches which enable sense-making of user data (Section 2.1). Second, we discuss systems which help analysts understand and explore users and groups using visual variables (Sections 2.2 and 2.3, respectively). Third, we discuss underlying connections between group formation and exploration, and review systems that implement partial handshaking (Section 2.4). Last, we discuss how current systems support alternating between the user and group context (Section 2.5).

2.1. Visual analytics for understanding user data

Visualization refers to a set of approaches which enable sense-making of data using visual variables [17]. It adds value to insights with the use of visual views rather than textual or tabular content [18]. The combination of analytical reasoning and visualization gave birth to the field of Visual Analytics (VA) [19, 20]. This field is responsible for the formulation, refinement, and validation of hypotheses about data using interactive visual interfaces. VUGA implements visual analytics for analyzing user data. A common challenge for visualizing user data is *clutteredness*, *i.e.,*, the huge volume and heterogeneity of user data hinders its effective visualization. We review related work which tackled this challenge to provide a clear visualization of user data.

Scatter plots and parallel coordinates. A preliminary solution is to employ visual views which can separate user data naturally, *i.e.,* scatter plots [21] and parallel coordinates [6]. However, the drawback of such views is that they only serve numerical attributes. Also, they do not fully solve the problem of clutteredness, as a rich user data with many attributes is still problematic to be

visualized with such methods.

Summarization. Another way to address clutteredness is to visualize only a summary of user data. In [22], a visualization approach is proposed to summarize sequences of user events and provide details-on-demand only. In [13], the analysis is limited to a pre-defined set of groups (*i.e.,* cubes) to reduce the amount of visualized data in each analysis iteration. However, the challenge with most summarization methods is that they are lossy, *i.e.,* there is no way to revert to the unsummarized version of the data, or it is time- and space-consuming.

Customizability. There is a recent trend whose effort is to provide a set of "visual grammars" where analysts can customize the analysis process and define what they want to see, hence reducing unnecessary content to visualize [23]. Vega is among the most popular visual grammars in the literature [10], where analysts can employ "signals" to associate their customized way of visualizing data to visual variables. Signals are dynamic variables that parameterize a visual element (e.g., a circle representing a group) for interactive behavior. Full customizability of visual grammars enables analysts to express clear visualizations for their analysis tasks. While visual grammars are beneficial to define the structure of an analysis task and reduce visual content, it is time-consuming, and analysts are not necessarily knowledgeable about visual grammar rules.

Dimensionality reduction. Recently, dimensionality reduction has become the method of choice in visual analytics to represent a clear 2D visualization of user data [24]. The proximity in the 2D view reveals the similarity between users. Popular dimensionality reduction methods are Principle Component Analysis (PCA), Multidimensional Scaling (MS) and t-distributed Stochastic Neighbor Embedding (t-SNE) [15], and more recently UMAP [25]. The focus of PCA is to capture variance in user data [26]. Given a user and its attributes, PCA uses its covariance matrix to perform a linear transform from the attributes to two new orthogonal dimensions with the largest possible variance (aka, the Rayleigh quotient). However, the linearity of PCA dismisses the similarities between group members. MS focuses on finding a matching from the n-dimensional space to a

2-dimensional space which preserves similarities between group members [27]. The advantage of MS over PCA is in its extended functionality to non-linear mappings. MS minimizes a stress function which captures the difference of user similarities between the original view and the 2D view. While t-SNE has the same manifold nature as MS, it focuses on local structures of group members to obtain a clearer view [28]. Instead of the stress function, t-SNE minimizes the KL-divergence between the distribution of user similarities in the original view and the 2D view to separate dissimilar members even more. For all these reasons, we employ t-SNE in this work to obtain a clear 2D view of user data. UMAP [25] is the newest among these dimensionality reduction techniques. It has similar visualization capabilities compared to t-SNE, but at a lower cost.

2.2. Visual enablers for group formation

The integration of VA approaches with group formation enables visual inspection of user groups. Analysts inspect formations in a visual form, and if they are not satisfied enough for their task, they change filters to form other groups. Self-Organizing Maps [29] are employed to visualize the overall distribution of events in user groups. Belt charts (or Sunbursts) [30] are also used to provide a more focused view on biases in distributions (dominating attribute-values) of a single group (*e.g.,* presence of more females in groups than males.) In case the analyst wants to focus on one specific facet of formed groups, a 3D regression heatmap can be adapted to user groups to organize all groups in a 3D grid reflecting the extent of correlation between groups and the given facet [31]. In the work of Makanju *et al.* [32], hierarchical relations between groups are visualized to provide a big picture of the group space. PivotSlice focuses on one single group and visualizes relations between users in that group [33]. TruGRC [34] describes a group recommendation system that relies on aggregation strategies of user profiles. In VUGA, analysts observe users and their similarities in a 2D view. Selecting each subset of users in this view renders immediately a set of statistics about them, which help analysts decide which users to consider for group formation.

2.3. Visual enablers for group exploration

The integration of VA approaches with group exploration enables interactive visualizations for user groups. Visualized groups facilitate expressing needs for analysts [7, 35]. Zenvisage provides a visualization view for query-based explorations [7]. Data Tweening is another visualization view which enables analysts to retain changes between consecutive iterations of exploration [8]. Vexus is a visualization framework to provide native support for exploring user groups [9]. FlashView is also a visualization interface for fast exploration of user groups using approximate query processing (AQP). In VUGA, a seed group will be selected in the 2D view and the system will explore other similar and dissimilar groups to the seed group.

2.4. Handshaking between group formation and exploration

Group formation and exploration have each their separate systems and it is often hard to enable a natural *formation*⟳*exploration* loop, due to their different nature. Few approaches in the related work incorporate them into the same loop. In [11, 36], groups are formed as frequent patterns, and the analyst can only examine groups selected by an objective function. In [12], the focus is more on exploration where groups are sampled from the group space and analysts can guide the sampling process according to their interests. Also in [13, 14], a set of pre-computed groups are available for exploration. Boratto et al. [37] describe an approach for automatic group detection that follows a group modeling strategy. The main challenge in all such systems is that the interaction between formation and exploration is weak as one needs to start a group formation process from scratch when alternating from an exploration process.

2.5. Alternation between users and groups

Most user analytics approaches provide results either in the form of individual users or groups, hence there is no alternation between users and groups. In [29, 30, 31, 11, 36], user data is analyzed only in the form of groups, hence inquiring about their members is nearly infeasible. While groups and their members are

both considered in [33, 38], interactions with groups is missing. It is crucial for a user analytics system to provide means to alternate between groups and their users anytime during the exploration. In VUGA, analysts can select any subset of users from the 2D view to obtaining detailed information at the individual level. They can then handpick a few users among them for group formation.

3. Design Considerations

We discuss four design considerations underlying the development of a visual analytics system for user data and user groups.

3.1. Represent and visualize user demographics and rating data Rating user data involves demographic information about the users and the items they review. The visualization of demographics using charts offers a simple way to drive exploration for each of the demographic categories. A more challenging task is the creation of a visualization that allows exploration of users with similar interests (e.g., similar genres of rating items). Such a task requires the creation of a view of the data that allows analysts to inspect users with similar interests. VUGA represents user interests as feature vectors and maps each vector using a 2D projection. The 2D projection positions users based on their similarity offering an additional way to explore users in the subsequent steps (group formation and exploration).

3.2. Enable filtering and group formation

Given a set of users, the analyst should be able to filter them based on their demographics and interests. The system should also allow the analyst to build a seed group that will serve as a basis for further exploration. VUGA associates a set of statistics to the displayed users. Analysts can select users of interest using one of two ways: by simply using a lasso tool to select users, or by using demographics and interests filtering. In all cases, a seed group is constructed, and a coordinated view of various statistics associated with the seed group is provided. The coordinated view is updated as group membership evolves.

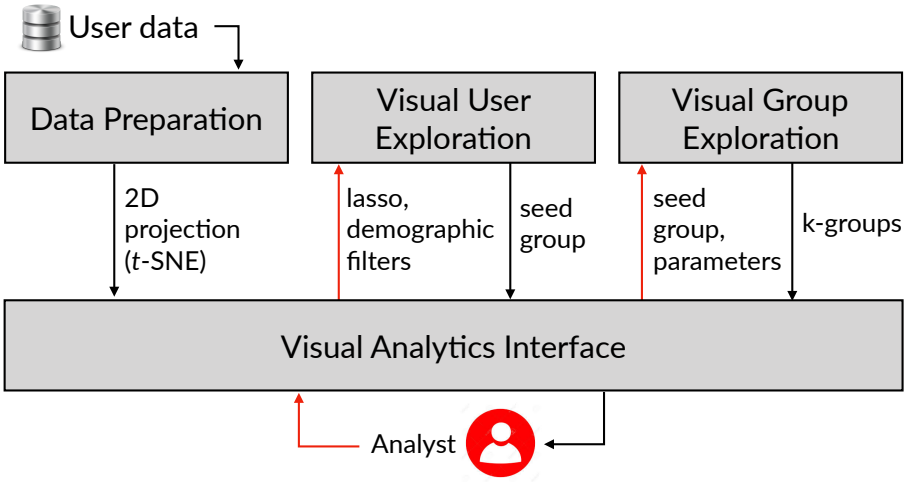


Figure 1: VUGA components. The red arrows show inputs from the analyst and the black arrows reflect the communication among components.

3.3. Enable group exploration

Given a seed group, the system must provide a way to use it as a starting point in exploring the space of existing groups. The exploration should not confine the analyst to specific regions of the space. In VUGA, the analyst can ask for similar/dissimilar groups. With this feature, the analyst can choose to maintain her train of thought by continuing to explore groups with similar demographics and interests or could choose to jump in the space of groups via the dissimilarity feature.

3.4. Enable handshaking between user data exploration and group exploration

This handshaking closes the loop between the two key components of user group analytics: analytics on users and analytics on groups. Analysts must be able to switch from a user-centric view to a group-centric view and vice versa. In both cases, VUGA relies on a data model that enables the efficient ingestion of user data and provides a visual interface to switch from user data to groups and vice versa.

4. The Components of VugA

VUGA[1] provides a visual analytics interface that communicates with three components depicted in Figure 1: the *Data Preparation* component that admits raw user data and produces a 2D projection that is made available to the analyst; the *Visual User Exploration* component that takes as input a 2D projection produced by data preparation and possibly filters specified by the analyst using a lasso tool or demographic attributes, and generates a seed group; the *Visual Group Exploration* component that admits a seed group and finds k similar/dissimilar groups (k is provided by the analyst). The visual analytics interface acts as an enabler to the components of VUGA and allows seamless integration of user exploration and group exploration. It also enables a full loop by allowing the analyst to select one or several groups out of the k groups returned by exploration as an input to the user exploration component to enable the inspection of their members and their statistics. We now describe the components of VUGA in more detail. First, we discuss the Data Preparation component which renders the visual layout (Section 4.1). Then we provide an overview of the visual analytics interface (Section 4.2). Last we present User Exploration and Group Exploration components (Sections 4.3 and 4.4).

4.1. Data Preparation

VUGA supports user data from different domains. For this purpose, it uses a generic format that allows direct integration with the visual interface. The format consists of four different tables: users U, items I, events E, and similarity features X. The table U contains user demographics (*e.g.,* age, gender, occupation, *etc.*), I contains descriptions of items (*e.g.,* movie or book information, medical treatments, *etc.*), E describes relations between users and items (*e.g.,* review of a movie or book from a given user, a medical treatment for a given user, *etc.*), and X describes features derived from user demographics. We represent derived data in a *similarity feature space* that contains an n-dimensional

[1] `https://github.com/FabianColque/VUGA`

record (where n is the number of genres) for each user, where similarity is defined as the distance between points in this space. We provide an example to explain how the similarity features are derived. Consider a dataset of movie reviews, and two users A and B, both having reviewed 100 movies. Suppose user u_1 reviewed 60 drama, 20 comedy, 15 romance, and 5 children's movies. Similarly, user u_2 reviewed 10 drama, 50 comedy, 25 romance, and 15 children's movies. One way to compare users is to encode the genre of movies as a 4D record using the percentage of reviews per genre. In this case, assuming that the dimensions are drama, comedy, romance, and children's, the result record for u_1 would be $\langle 0.6, 0.2, 0.15, 0.05 \rangle$ and for u_2 $\langle 0.1, 0.5, 0.25, 0.15 \rangle$. User similarity can then be computed using any metric in the feature space, such as Cosine and Pearson Correlation.

Analysts have access to both demographics and similarities through the interface. However, instead of interacting with the similarity space in a complicated n-dimensional space, we project it into two dimensions using a dimensionality reduction technique called t-SNE [15], which is widely applied in areas such as machine learning, visual analytics, and data mining. Like other dimensionality reduction techniques, t-SNE positions points closer to each other in the 2D projected space if their counterparts in the original n-dimensional space are also close. Many aspects impact the position of projected high-dimensional points. Ideally, projected points are the least overlapping possible while keeping similar points close to each other. t-SNE admits as input the number of iterations and perplexity. These two parameters dictate how the projection shapes up. In this paper, we computed projections in trial-and-error iterations of varying parameter combinations to choose the one with a reduced projection error and a good separation of points (*i.e.,* uncluttered view). Since the t-SNE computations are costly, the projection is performed offline by the Data Preparation component.

Although extremely powerful, the effective use of t-SNE is challenging. The guidelines proposed by Wattenberg and colleagues [39] suggest running several tests with varying values for t-SNE parameters. The first parameter is called *perplexity*, which roughly speaking, encodes a distance from a point to its

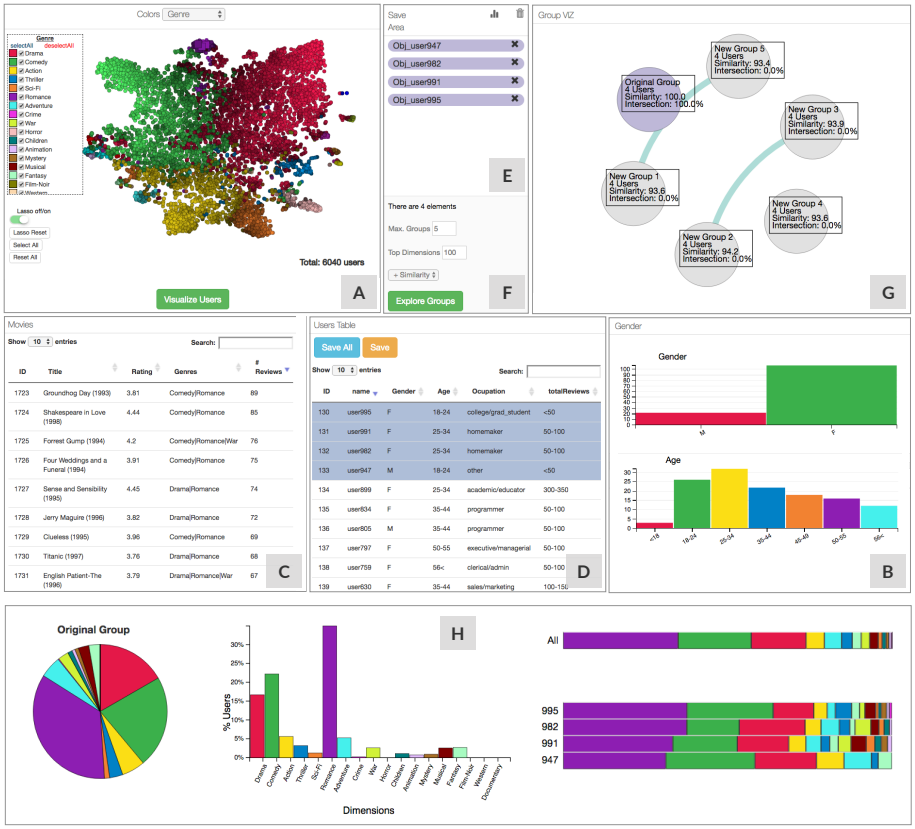
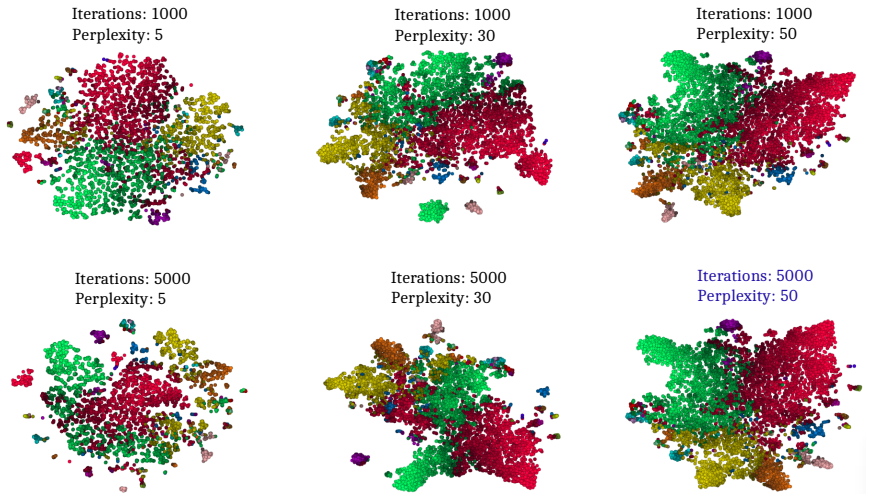


Figure 2: VUGA visual interface displaying the MOVIELENS dataset. User selection can be done either in can be selected either in view A or B. Other views are Other areas include: items view (C), user description view (D), user selection area (E), group exploration configuration (F), group exploration results (G), and group charts (H).

neighbors that aims to balance between local and global aspects of the data. The second parameter is the *number of iterations*, which indicates the maximum number of times an optimization process runs until a cost function C can be found optimal and adequate. According to experiments done in previous works [15], the recommended values for the number of iterations are 1000 and 5000 iterations, while for perplexity, the recommended values are 5, 30, and 50.

We ran several tests using a combination of these parameters and list the final cost function minimized during the optimization. In Table 1, we display the

Figure 3: Projection results using different $t-$SNE parameters in the MovieLens dataset.



resulting cost value for both datasets using a combination of perplexity (5, 30, and 50) and the number of iterations (1000 and 5000). Each combination was executed three times to account for variability in the results. We highlight for each dataset the minimum cost function value found. In the MovieLens dataset, the optimal values correspond to the perplexity of 50 and 5000 iterations. In Fig. 3, we display the projection obtained for varying values of perplexity and the number of iterations in the MovieLens dataset. By comparing the distinct projections, we observe that the projection using the optimal parameters found in our tests (iterations = 5000 and perplexity = 50) created a projection that does a better job in separating the users with a strong preference for the three main genre classes (drama-red, comedy-green and action-yellow.

4.2. Visual Analytics Interface

Figure 2 illustrates the visual analytics interface of VugA. The interface has distinct views to display information and statistics of users, configure group formation and exploration, and display the generated groups. VugA uses a coordinated user interface that updates the information displayed after any

Table 1: Varying $t - SNE$ hyperparameters (perplexity and nunber of iterations) for the MovieLens and BookCrossing datasets. Each configuration was executed three times. The configuration with the minimum cost function value is chosen to generate the projection.

Configuration			Cost	
# Run	# Iterations	Perplexity	MovieLens	BookCrossing
1	1000	5	1.802657	1.199160
2	1000	5	1.765758	1.203062
3	1000	5	1.778850	1.193867
1	5000	5	1.575450	1.026430
2	5000	5	1.556541	1.027299
3	5000	5	1.613056	1.025825
1	1000	30	1.693772	1.122926
2	1000	30	1.705552	1.134787
3	1000	30	1.673824	1.135139
1	5000	30	1.649885	1.097255
2	5000	30	1.672916	1.095400
3	5000	30	1.669642	1.087319
1	1000	50	1.570772	1.070013
2	1000	50	1.578098	1.077052
3	1000	50	1.5696902	1.074400
1	5000	50	1.5583727	1.051515
2	5000	50	1.5557487	1.040932
3	5000	50	1.5586364	1.043676

selection. Different views of this interface are explained as follows.

User projection view (A) displays a collection of points (from the tables X and E of user data), each corresponding to a user. The t-SNE projection defines the position of users. We color-code points using a set of pre-defined mappings. Users can be selected in this view using a lasso tool.

Demographics view (B) displays statistics over the demographic attributes (from the table U of user data). Users can be selected directly over each demographic attribute, and the coordinated interface gets updated automatically right after each analysis iteration.

Items view (C) lists items (from the table I of user data) associated with the

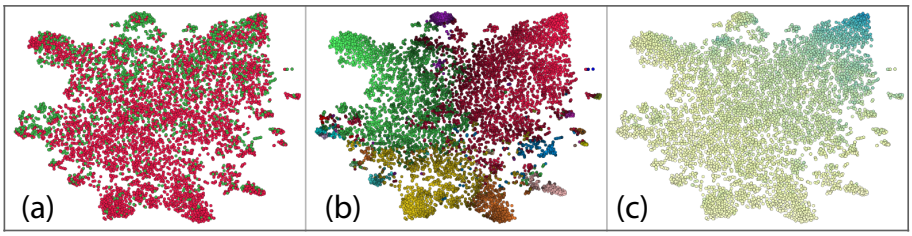


Figure 4: Variants of color mapping based on (a) gender (categorical scale), (b) dominant genre (categorical scale), (c) degree of genre dominance (continuous scale).

current selection of users. Items can be sorted according to any column in the table.

User description view (D) lists user's detailed information (from the table U of user data) in a table. Users can be sorted according to any column in the table.

User selection area (E) is a buffer that lists the users saved in the group formation phase (by the analyst) and which constitute the seed group for group exploration.

Group exploration configuration (F) is an area for configuring the group exploration parameters, including objective (similarity or dissimilarity) and the number of groups to return (k).

Group exploration results (G) is an area for displaying the results of group exploration.

Group charts (H) displays statistics about the seed and explored groups in the form of pie charts, bar charts, and stacked-bars.

The user projection view enables analysts to visually inspect similarities between users. Note that there is no notion of coordinate axes in the 2D projection, and the view reflects only the proximities. However, other visual variables (*i.e.,* color hue, brightness, transparency, and glyphs) are employed in the 2D view to providing more information about users. For instance, we use different color mappings to differentiate among users attributes. Analysts can

configure the interface to define the mapping between attributes and colors, color scales, and weights of color brightness. Figure 4 shows different configuration settings, where each point depicts a user. The example on the left shows a categorical red-green color-coding, where gender is used for mapping. The middle example maps a categorical color table to each genre. The color of each point is defined by the dominant genre of its corresponding user. The dominant genre of a user is the one whose number of reviews is the largest. The color weight of the point is also defined by the amount of dominance. For instance, if Drama is the dominant genre of a user covering 80% of her reviews, a bright red point (red is associated to Drama) represents this user in the 2D view. The example on the right features a continues color-coding mapped to the amount of genre dominance. Such color-coded views enable the comparison of all genres in one place. We consider the middle example of Figure 4 as our default view in VUGA's interface.

4.3. Visual User Exploration

The input to User Exploration is the 2D projection of user data. This component supports the ability to explore statistics about users and build a seed group out of individual users. This is done in two steps: *selection* and *inspection*. First, the analyst picks a subset of users either by selecting demographics or by using a lasso tool. Then she inspects the information and statistics provided for these users to handpick few users of interest, which constitute her seed group for further exploration.

User selection. Analysts can use two different views of VUGA's interface to select users or refine a previous selection: user projection view and demographics view, *i.e.,* views A and B in Figure 2, respectively. Both views are coordinated, hence a selection on one reflects the other. Figures 5-A and 5-B illustrate the selection process using the user projection area. First, a lasso tool is provided where analysts can draw freehand selections around the users in the projection. There is also a categorical selection which allows the analyst to choose dominating genres of interest. Figure 5-C illustrates the selection process

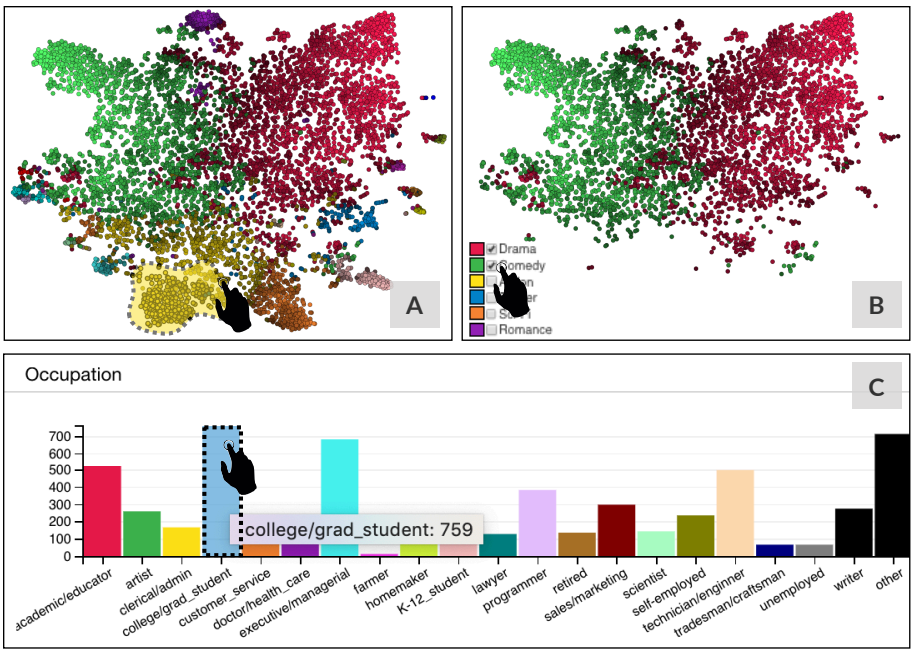


Figure 5: Selection methods for group formation: lasso selection in the user projection area (A), category selection in the user projection area (B), and demographics selection (C).

in demographics view, where attributes can be selected by clicking on their corresponding histogram bar (*e.g.,* filtering users to college/grad students in the figure). The selection process is composable, *i.e.,* a later selection can add/remove users to/from an initial selection. For instance, a lasso selection can then be refined by selecting a few demographic attributes of interest.

User inspection. In Figure 2, views C and D are used for inspection. They list the items that the selected users are associated to (*i.e.,* reviewed movies/books), and demographics of those users, respectively. The analyst can handpick users of interest in the user description view (Figure 2-D). Handpicked users will be added to the user selection area (Figure 2-E). Group charts (Figure 2-H) displays some aggregated statistics (in the form of pie charts, bar charts, and stacked-bar charts) for the current selection of users in the user selection area. This helps the analyst to see how insightful her current selection is. While the user

projection view (Figure 2-A) only convey genre dominance information, group charts provide more details in the form of distributions of different demographic attributes.

4.4. Visual Group Exploration

Group exploration admits as input a seed group returned by group formation and generates similar or dissimilar groups. Few parameters dictate the functionality of group exploration. First, we review these configurable parameters, and then we discuss the process of group exploration.

Group exploration parameters. Analysts can configure the parameters of group exploration in the "group exploration configuration" view (Figure 2-F). The first parameter, k, is the number of groups to return by the group exploration process. The second parameter (*i.e.,* top dimensions) allows discarding negligible dimensions. The parameter specifies the percentage of the top most relevant dimensions of the seed group. For instance, if the selected users in the user selection area expressed reviews on 18 different genres, then tuning this parameter to 90% would consider 16 most relevant genres for exploration. The parameter is set to 100% by default (*i.e.,* use all dimensions). The third configuration parameter defines the objective of exploration, be it either similarity or dissimilarity. For instance, if the analyst sets $k = 5$, top dimensions to 100% and similarity as the objective, then 5 similar groups will be generated as exploration options using all available dimensions of the seed group.

Group exploration process. Given a seed group, group exploration returns k similar/dissimilar groups and illustrate them in the "group exploration results" view (Figure 2-G). Each new group can become a seed group in the next analysis iteration. Figure 6 illustrates the process of group exploration in VUGA. For each user in the seed group, we compute all neighbor users in a radius r and then select the top-m closest/farthest neighbors, in case the objective is similarity/dissimilarity, respectively. Note that the similarity computation and comparison are performed in the original n-dimensional space. Then for each

user u, we obtain a list N_u which contains m neighbor users of u at the distance r. A candidate group can be generated for exploration by randomly selecting k users from neighbor lists. Other candidate groups will be generated in the same fashion but with no user intersection with previously generated groups. The parameters r and m define the "pool size" and "heterogeneity" of exploration, respectively. The higher r and m are, new groups pick more heterogeneous users from a larger pool. Note that these parameters are not explicitly placed in the interface. However, analysts can tune these parameters in the back-end configurations. In our use cases, we set r to the 10% of the largest distance between a pair of users and $m = 50$ as a result of minimizing error in a k-fold cross-validation.

VugA is generic and can incorporate different group exploration methods. In this paper, we focused on the similarity/dissimilarity exploration method as an intuitive way of sensemaking in user data. However, other methods can also be integrated in VugA, e.g., diversity exploration [14], contrast group exploration [40], distribution exploration [41], and multi-objective exploration [1]. One of our future directions is to perform a thorough study on several exploration methods and their influence on user group analysis.

5. VugA in Practice

In this section, we demonstrate usage scenarios of VugA in exploring users and exploring user groups in MovieLens and BookCrossing datasets. To focus on the practical aspects of our approach, we provide use cases for analysts in their role as *domain experts* (Section 5.1) and *novices* (Section 5.2). In each case, we define a hypothetical exploratory scenario and describe the exploration steps and the results obtained. In Figure 7, we show dominant genres for each dataset in the projection to illustrate areas where a given genre dominates others. We used different sets of t-SNE parameters for each dataset to obtain the most uncluttered projection. For MovieLens, we use 5000 iterations and perplexity of 50. For BookCrossing, we use 1000 iterations and a low perplexity of

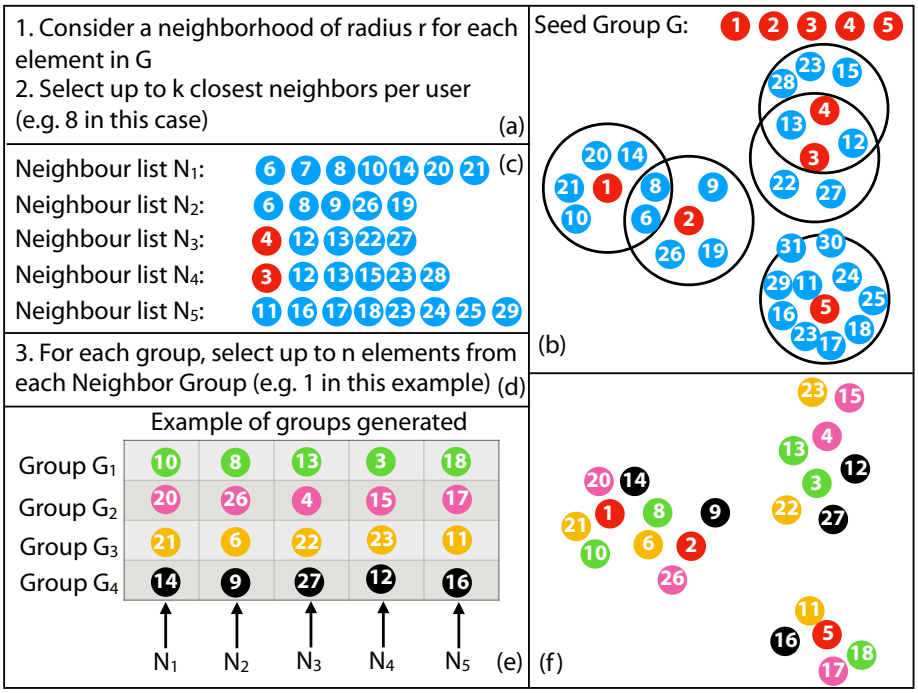


Figure 6: Illustration of the group exploration process. In this example, four new similar groups are generated from a seed of group of five users (users 1 to 5).

5. A glance at the figures shows that Drama and Comedy are dominant in MovieLens, while Non-Fiction and Mystery are the most reviewed genres in BookCrossing.

5.1. Finding Movie Critics

MovieLens is a movie review dataset[2] consisting of over 1M ratings for 3,952 movies given by 6,040 users. User attributes are ID, name, gender, age, and occupation. Movie attributes are ID, title, and the list of genres for the movie. The ratings relate a user to a movie and contain a score (from 0 to 5) which reflects the user's opinion on the movie. As instructed in Section 4.1, first we build a feature space which encodes dominant genres. Since there are 18

[2] *https://grouplens.org/datasets/movielens/1m/*

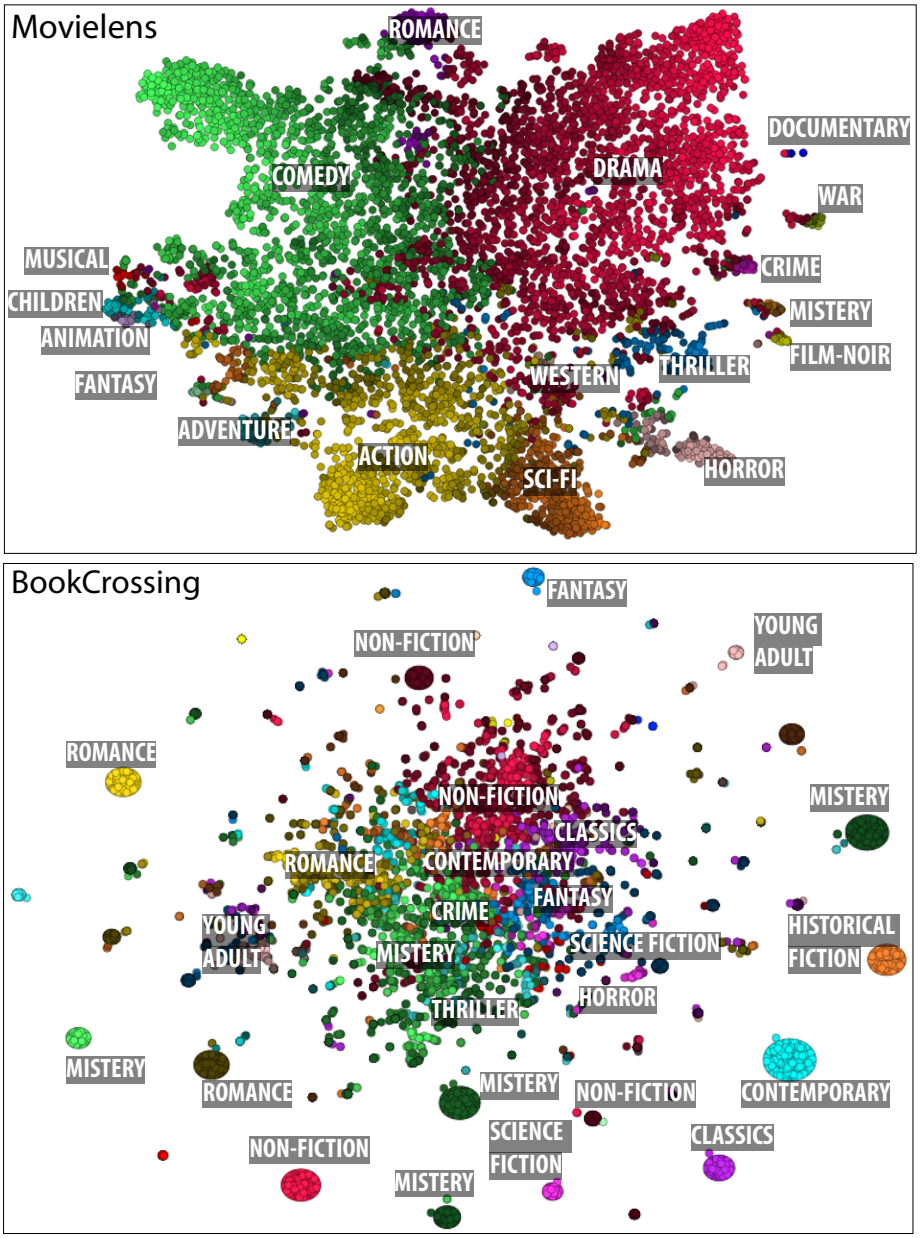


Figure 7: Projection view for MovieLens and BookCrossing with dominant genres mapped over distinct regions.

different genres in MovieLens, the feature vectors have 18 dimensions, each normalized by the total number of reviews.

Exploration task. Our goal in this use case is to gather a diverse set of reviews at the first screening of Drama and Comedy movies. We set out to identify a group of reviewers to achieve that goal. The generated group should contain 5 reviewers with Comedy as the dominant genre, another 5 reviewers whose dominant genre is Drama, and a mix of 10 additional reviewers who differ from them either in demographics or in interests. In the following, first we describe the role of the analyst as a domain expert, and then illustrate how exploring users and groups in VugA enables the analyst to find a set of movie reviewers.

Exploration as a domain expert. In this use case, we consider a domain expert who has a rich knowledge about the exploration task, and also the attributes and the structure of the underlying data (i.e., MovieLens). However, it is often the case that the background knowledge is only subjectively available in the mind of the analyst, and is not formally represented in form a knowledge base. Hence it is almost infeasible for such analyst to directly query or filter out data based on her needs. In this case, the analyst requires to interact with the system in several iterations to describe her needs and obtain relevant results. It is also crucial for domain experts to access details of their data in different levels of granularity, and the system should provide enough support to navigate between those different levels.

User exploration. We illustrate the exploration process in Figure 8. The domain expert starts the exploration by looking for reviewers that have Comedy as their dominant genre. For this purpose, she selects reviewers with Comedy as their dominant dimension using the categorical checkbox in the projection view. This selection limits the view to only green shaded points (the color green is associated with the Comedy dimension). Then she employs the lasso tool to select a set of reviewers in the top-left area of the projection. The result of this selection is 251 reviewers, and the coordinated interface is updated to only display the information associated with the current selection. The expert

performs subsequent selections in the demographics to narrow down the age range to 25-34 years old, resulting in 78 reviewers (44 male and 34 female). She also selects the occupation as academic/educator, resulting in a group of 11 reviewers (6 male, 5 female). She inspects the stacked-bar histogram for those 11 reviewers, which are ordered top-down by the dominance factor, and saves the top-5 reviewers in the user selection area (2 male and 3 female). The analyst repeats the process above after changing the dominant genre to Drama. The result of this process generates an additional 5 reviewers, leading to a seed group of 10 reviewers (6 male and 4 female).

Group exploration. The exploration task is to find a mix of 5 to 10 reviewers who differ from the seed group in demographics and interests. Given the seed group, the domain expert invokes the group exploration algorithm to generate the 3 most similar and the 3 most dissimilar groups. All generated groups are distinct from the seed group. By a visual comparison of the stacked bar chart of the seed group against all 3 most similar groups, she observes that the histograms are consistently similar, which is also confirmed by computing the Kullback-Leibler divergence. Also, 2 out of 3 most similar groups have Drama, Comedy, and Romance (in this order) as dominant genres, while one group has Comedy, Drama, and Romance. The expert examines the demographics of reviewers in those 3 groups and handpicks 5 reviewers whose age > 34 and whose occupation differs from academic/educator. She adds those 5 reviewers to the seed group resulting in a total of 15 reviewers who have similar dominant genres and different demographics. Similarly, she examines the most dissimilar groups to the seed group and identify Horror, Thriller, and Sci-Fi as their dominant genres. She handpicks 5 reviewers whose demographics are age within 25-34 and occupation is academic/educator. By adding those 5 reviewers to the seed group, the analyst ends up with 20 reviewers in total.

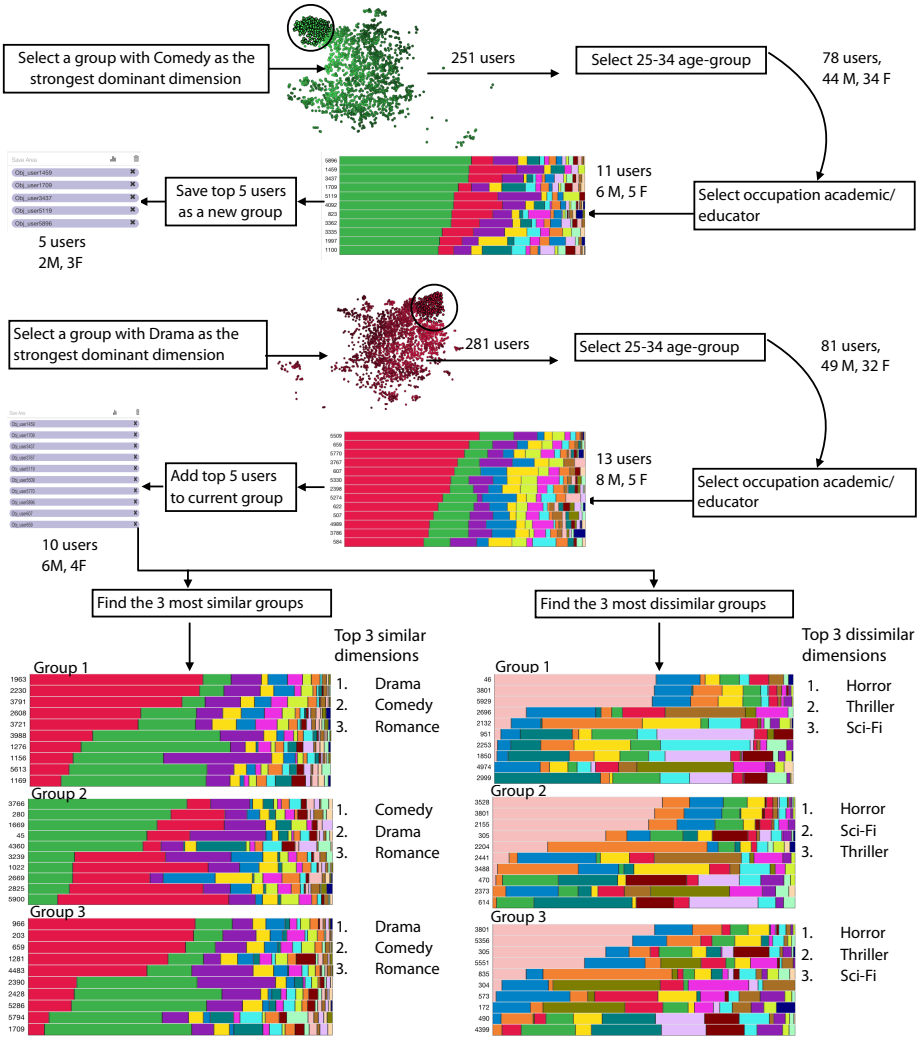


Figure 8: Exploring the MovieLens dataset to gather a diverse set of reviews at the first screening of Drama and Comedy movies.

5.2. Forming a Book Club

BookCrossing is a book review dataset[3] with 101,376 ratings for 46,380 books given by 8,167 users. Users have a unique ID, a name, a location, and age,

[3] *http://www2.informatik.uni-freiburg.de/~cziegler/BX/*

while books have an ID, a title, an author, a year of publication, a publisher and a genre. Since the original BookCrossing dataset does not contain information on book genres, we append the genre information crawled from the Goodreads website[4]. Ratings relate a user to a movie (in a scale from 1 to 10). We built a feature space that encodes dominant book genres (in the same way we did for MovieLens).

Exploration task. Our goal in this use case is to build a book club for Romance that appeals to senior readers, to delay their memory decline [42]. We set out to identify the most popular authors among readers in that age range. We also look for books that received a high number of reviews among similar readers. In the following, first we describe the role of the analyst as a novice, and then illustrate how exploring users and groups in VugA enables an agnostic analyst to build a book club.

Exploration as a novice. In this use case, we consider that the analyst does not have a rich knowledge about the task and the data structure, and she builds her knowledge by observing insights on users and groups. As the analyst's knowledge is built iteratively, it is of critical importance that the analyst receives the most interesting set of insights at each iteration. In other words, VugA should act as a "guidance mechanism" which helps the agnostic analyst make wiser decisions.

User exploration. We illustrate the exploration process in Figure 9. The novice analyst starts by looking at the projection to obtain a bird's-eye view of the data. Points colored in yellow have Romance as their dominant genre. To clean the projection view, she selects Romance, which filters the view to only show relevant points. Given the high number of points, she employs the lasso tool to select a subset of those points. Then the analyst refines her search to only consider people above 54 years old. The result is a group of 43 reviewers who have as dominant genres Romance, Contemporary, and Historical Fiction.

[4] *http://goodreads.com*

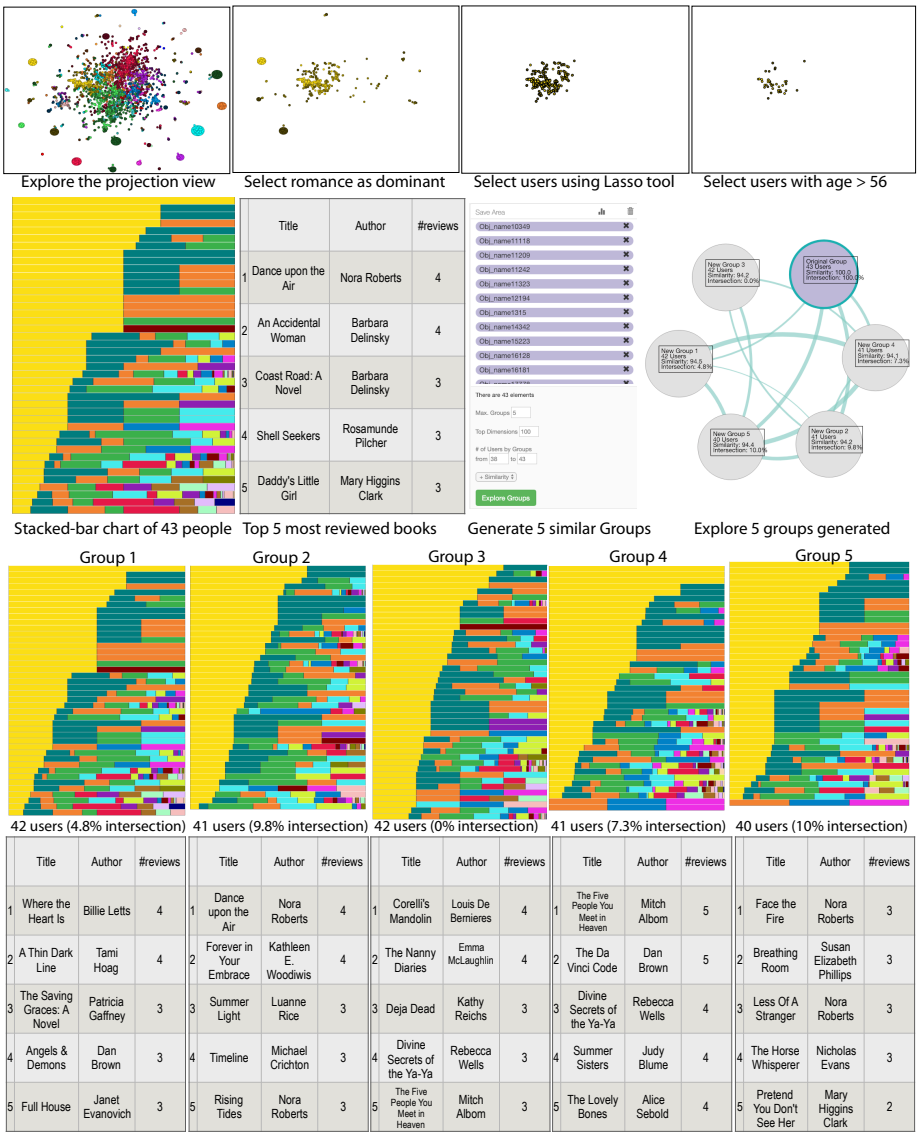
She examines the contents of that group and notices that only 17 books out of 2,657 received more than 3 reviews! Since we are interested in popular authors, the analyst selects the top-5 most reviewed books only and find that *Barbara Delinsky* has two books on that list. She marks her as an author of interest.

Group exploration. Then the novice attempts to find additional popular authors of Romance books. To do that, she saves the group of 43 reviewers as a seed group and asks the algorithm to generate 5 similar groups, resulting in groups that contain about 40 reviewers each. The analyst examine the contents of those groups carefully and find that one has no reviewers in common with the seed group and, 4 have up to a 10% overlap in reviewers. The analyst observes that the author *Nora Roberts*, who was present in the seed group, also appears in the generated groups four times with different books. That leads us to mark her as an author of interest. Additional authors, not in the seed group, like *Dan Brown*, *Mitch Alborn*, and *Rebecca Wells* appear twice. She also marks them as authors of interest. The analyst observes additional two books : *Divine Secrets of the Ya-Ya Sisterhood* that has received a total of 7 reviews across all groups and *The Five People You Meet in Heaven* with a total of 8 reviews across all groups. She marks those as books of interest.

6. User Study

We performed a within-subject user study [43] to evaluate the usefulness of VugA in practice using the MovieLens dataset. We recruited 16 subjects each of which took the role of an analyst in the study. Initially, subjects were given an interactive tour of our tool (one minute, approximately). Then, they were asked to complete five tasks which consisted in exploring the dataset and answering questions about the results they obtained. Subjects were also asked whether they encountered any problems during the execution of the task and had the opportunity of giving feedback. At the end, they were asked to complete a usability questionnaire [44].

Most subjects were male and were between 21 to 30 years old. To neutralize



	Title	Author	#reviews
1	Dance upon the Air	Nora Roberts	4
2	An Accidental Woman	Barbara Delinsky	4
3	Coast Road: A Novel	Barbara Delinsky	3
4	Shell Seekers	Rosamunde Pilcher	3
5	Daddy's Little Girl	Mary Higgins Clark	3

	Title	Author	#reviews
1	Where the Heart Is	Billie Letts	4
2	A Thin Dark Line	Tami Hoag	4
3	The Saving Graces: A Novel	Patricia Gaffney	3
4	Angels & Demons	Dan Brown	3
5	Full House	Janet Evanovich	3

	Title	Author	#reviews
1	Dance upon the Air	Nora Roberts	4
2	Forever in Your Embrace	Kathleen E. Woodiwis	4
3	Summer Light	Luanne Rice	3
4	Timeline	Michael Crichton	3
5	Rising Tides	Nora Roberts	3

	Title	Author	#reviews
1	Corelli's Mandolin	Louis De Bernieres	4
2	The Nanny Diaries	Emma McLaughlin	4
3	Deja Dead	Kathy Reichs	3
4	Divine Secrets of the Ya-Ya	Rebecca Wells	3
5	The Five People You Meet in Heaven	Mitch Albom	3

	Title	Author	#reviews
1	The Five People You Meet in Heaven	Mitch Albom	5
2	The Da Vinci Code	Dan Brown	5
3	Divine Secrets of the Ya-Ya	Rebecca Wells	4
4	Summer Sisters	Judy Blume	4
5	The Lovely Bones	Alice Sebold	4

	Title	Author	#reviews
1	Face the Fire	Nora Roberts	3
2	Breathing Room	Susan Elizabeth Phillips	3
3	Less Of A Stranger	Nora Roberts	3
4	The Horse Whisperer	Nicholas Evans	3
5	Pretend You Don't See Her	Mary Higgins Clark	2

Figure 9: Exploring the BookCrossing dataset. In this case we explore a group of people that prefer Romance genres and are older than 54 years-old.

the impact of expertise on our user study, we assume that all our subjects are "partially informed users", where they don't necessarily have a full knowledge of the datasets and the tasks. However, we asked about their experience with

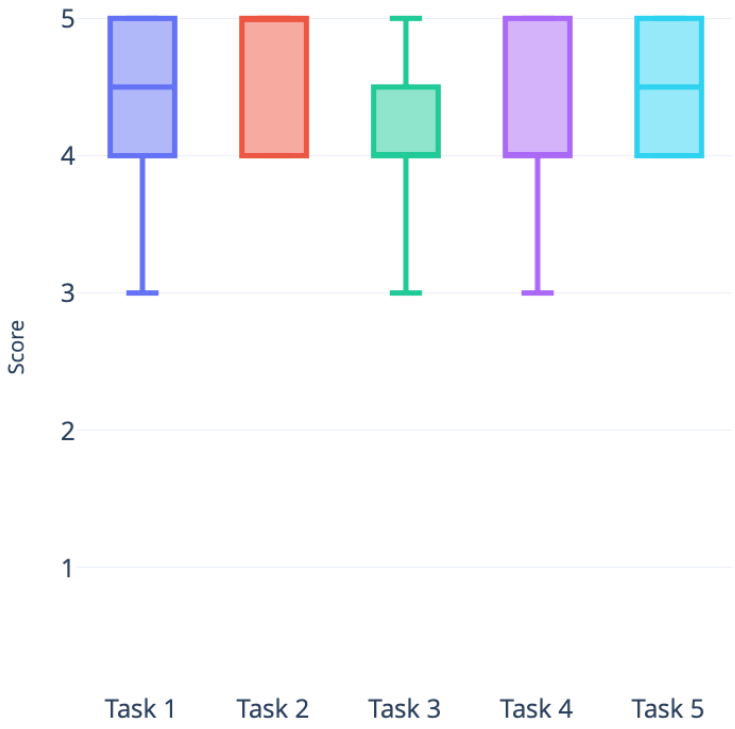


Figure 10: Boxplot of task questionnaire results using the Likert scale.

data visualization tools in pre-questionnaire, where we found that 44% have knowledge of info-graphics used in websites and newspapers, 68% use traditional graphics techniques at work, 19% use data visualization tools at work, and 50% use data visualization in their research projects (experience choices were not exclusive in our study). We defined five tasks including common actions (such as selecting, filtering, and saving users), followed by group exploration and analysis.

Based on the categorization of visualization tasks in [45], we considered the following list of tasks in increasing order of difficulty.

- **T1**: Using the lasso tool, select users who have a strong preference for Sci-Fi movies. What are the most common demographics (gender, age group, and occupation) in the selected group?
- **T2**: For each one of the most frequent demographics identified in T1, click

over the highest bar to narrow down the selection of users. Save the selected users in the Save Area. How many users have you obtained? What are the three most dominant genres in this group of users?

- **T3**: The list of saved users corresponds to your seed group. In the group exploration area, generate n groups of similar users. Inspect the new groups and choose one of the new groups. Find films that appear both in the seed group and in the new group chosen. Among those, identify the two movies with the highest number of reviews.

- **T4**: Using the groups generated in T3, explore the genre charts for the original and the new group. Movie genres in the stacked bar charts are presented in order of dominance. Are these stacked bar charts different across groups? If so, describe what you observed.

- **T5**: Generate groups that are most dissimilar in relation to the seed group. Explore these groups and identify the three dominant genres in the dissimilar groups.

Each task contains questions which help us assess whether the subject was able to achieve the intended goals of the exploration. Although some questions did not have an exact answer (as the answer depended on the actual selection of users made by the subject), we were able to have an inkling of the possible range of answers – we carried out the tasks ourselves a number of times and made notes about the possible answers.

For each task, we recorded the subjects' comfort in completing the task using a Likert scale, from 1 (difficult) to 5 (easy). With the user study, we can assess whether the design considerations discussed in Section 3 were achieved.

- Represent and visualize user demographics and rating data (Section 3.1) – This goal was measured in tasks Tasks T1 and T2. The questions in T1 were answered correctly by 15 out of 16 subjects. In T2, one of the subjects reported the total number of users rather than the number of selected users, and the same subject who had a problem understanding T1, also reported

movie genres that were not in the possible range of correct answers. Thus, for T2 we have 15 subjects with correct results. The successful completion of T1 and T2 showed that this design goal was met.

- Enable filtering and group formation (Section 3.2) – T3 assessed whether subjects' were able to perform these tasks in VUGA. Again, only one of the subjects reported implausible answers to the questions associated with T3. These results confirm that this design goal was also met.

- Enable group exploration (Sections 3.3 and 3.4) – Tasks T4 and T5 evaluated these design considerations. All answers to questions associated to these tasks were plausible, indicating that the visual comparison of groups using stacked bar-charts allowed the analysis of (dis)similarity across groups.

Figure 10 shows the boxplot of the task completion assessments of the subjects in the Likert scale. We observed that subjects felt comfortable in completing all five tasks, with a median value equal to or greater than 4. We ran an Analysis of Variance (ANOVA) to test whether there was a difference in mean across the five tasks. The calculated p-value was 0.068, which means that we accept the null hypothesis stating that there were no differences across tasks at a 95% confidence level. Since the tasks were in increasing order of difficulty, we attribute this result to two possible causes: the learning effect experienced by subjects as they progressed in the exploration tasks, or simply because all tasks were fairly easy for them to accomplish.

At the end of the tasks, users were asked to complete the System Usability Scale (SUS) questionnaire [44], which is a widely used tool for evaluating usability. VUGA scored 76.4 points. A study by Sauro *et al.* [46] analysing 500 SUS questionnaires concluded that a SUS score greater than 68 are above average and that a score of 74 converts at percentile rank of 70%. Thus, a score of 76.4 means that VUGA fared better than at least 70% of the systems analysed by the authors. Figure 11 shows a stacked bar chart of all 16 questionnaires completed by the subjects of the user study. For visualization purposes, agreement/disagreement with positive/negative questions on a scale from 1 to 5 were converted to a scale

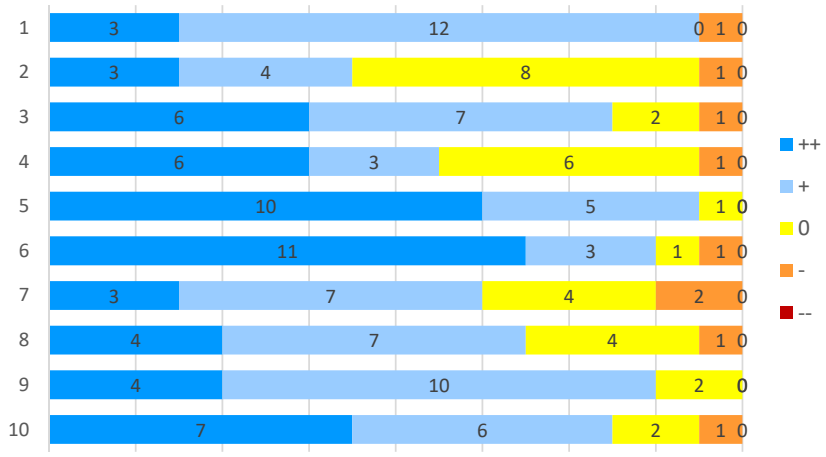


Figure 11: Results for System Usability Scale questionnaire.

from '- -' to '++', according to the opinion polarity they conveyed. No subject has given a strongly negative score to any of the ten questions. Out of the ten questions, the most positive opinions were regarding questions 5–"*I found the various functions in this system were well integrated*", for which there was a strong agreement and 6–"*I thought there was too much inconsistency in this system*", for which there was a strong disagreement. Most negative opinions were given in response to question 2–"*I found the system unnecessarily complex*", for which half the subjects chose the 'neutral' answer, one subject gave a negative, score and the remaining seven gave positive scores. We believe this happened because most visualization systems that the subjects are used to interacting are simpler than VUGA.

Limitations. Despite the positive results of the user study, the feedback given by some subjects have pointed out limitations of VUGA. Two subjects mentioned that the need for scrolling up and down a few times in order to complete tasks T4 and T5 made them lose focus. One subject suggested that movie genres should be listed alphabetically. Finally, one subject mentioned she would have liked to select multiple groups to compare (and not just two).

7. Summary and Discussion

We described VUGA, a visual enabler for user data exploration and group exploration. We provided representative tasks and showed how VUGA helps analysts achieve different goals. Despite its strong points, VUGA has limitations. Some subjects in the user study identified specific features they wish were adjusted. Furthermore, the number of colors that we can assign to the circles in the Projection Area is limited by the number of colors a human can perceptually distinguish. In this article, we used a maximum of 18 colors/categories and feel that this would upper bound – studies recommend using no more than ten colors [47]. In addition, the number of users and groups that can be clearly displayed depends on the size of the screen – *e.g.,* in a small laptop screen, a user can comfortably see at most ten groups. We designed our user interface to accommodate a small graph (with up to 20 groups) because the group exploration task we envisioned were also limited by a small number of groups the user wants to simultaneously compare. If necessary, increasing this limit would require scaling the graph drawing technique to support the display of larger graphs, as discussed in the work of Graves [48].

A user study is limited to qualitative analyst-centric evaluations. Hence the need for a principled evaluation methodology arises, which we consider as a future perspective. Despite the established body of related work for evaluating user data exploration, group exploration, and visualization alone, there is no evaluation methodology for their combination [4, 49, 50]. A valid question is whether *we can evaluate* VUGA *with a combination of methods proposed to evaluate its components?* Adapting exploration-based and visualization-based evaluation protocols (*i.e.,* quality and user experience axes) does not cover its quantitative aspects (*e.g.,* how fast its formation method performs.) While user studies alone are often biased and incomplete [51], we discuss four novel opportunities of all-together evaluation of user exploration, group exploration, and visualization, as follows.

Isolation. A natural approach is to isolate human-oriented aspects such as user

experience and satisfaction and evaluate other aspects (*i.e.,* performance and quality of results) using traditional measures. For human-oriented aspects, a user study is often designed. Although isolation enables a thorough evaluation, it suffers from two drawbacks. First, the boundaries of human-oriented and system-oriented aspects are fuzzy. For instance, group exploration is fired by an analyst, but some system-oriented aspects (*e.g.,* amount of intersection between groups) are also associated with exploration. Second, isolated evaluation assesses user exploration and group exploration separately, and does not capture their interactions.

Crowdsourcing. Crowdsourcing platforms such as Amazon Mechanical Turk[5], Crowd4U[6], and Figure Eight[7] scale up user studies by providing access to a large audience of information consumers [52]. The high confidence associated to a user study with a large population dissolves doubts on bias and incompletion. It is shown in [53] that for a dataset with more than $100K$ users, at least 1100 participants are needed to achieve results with an error margin of $\pm 3\%$.

Quantified user study. User studies can be enriched with quantified measures to complement participants' answers. While responding to questions, measures such as time-to-think, mouse actions, eye movements, scrolling actions, dragging speed, number of backtrackings, number of cycles, and number of restarts will be recorded for participants [54]. This enables both qualitative and quantitative evaluations.

Benchmarking. Quality can be assessed by comparing it against standard tests, *i.e.,* benchmarks. Benchmarks are a common practice in the database community (*e.g.,* Oracle TPC benchmark [55] and LDBC Social Network Benchmark [56]) In [57], a few visual exploration benchamrks are discussed, such as IDEBench [58] and REACT [59]. A benchmark should consist of analyst traces, *i.e.,* a recorded session (using screen captures, recorded voice, I/O capture, *etc.*) of analyst

[5] *https://www.mturk.com*
[6] *https://crowd4u.org*
[7] *https://www.figure-eight.com/*

actions in every component of the system they are interacting with.

In summary, the evaluation of an interactive visual analytics tool for exploring users and user groups would need to go far beyond typical user studies and quantitative measures. Appropriate benchmarks that capture human factors (*e.g.,* motivation and satisfaction) in user group exploration need to be designed and deployed. We believe that such an effort must be adapted to different application needs and hence result in domain-specific benchmarks.

Acknowledgements: This work was partially supported by CAPES-COFECUB, CNPq/Brazil, and the CDP LIFE C7H-ID16-PR4-LIFERH grant in Grenoble.

References

[1] B. Omidvar-Tehrani, S. Amer-Yahia, P. Dutot, D. Trystram, Multi-objective group discovery on the social web, in: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I, 2016, pp. 296–312.

[2] B. Omidvar-Tehrani, S. Amer-Yahia, R. M. Borromeo, User group analytics: Hypothesis generation and exploratory analysis of user data, VLDB Journal.

[3] F. Yang-Wallentin, K. G. Jöreskog, H. Luo, Confirmatory factor analysis of ordinal variables with misspecified models, Structural Equation Modeling 17 (3) (2010) 392–423.

[4] B. Omidvar-Tehrani, S. Amer-Yahia, User group analytics: Discovery, exploration and visualization, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018, 2018, pp. 2307–2308.

[5] K. Hu, D. Orghian, C. Hidalgo, Dive: A mixed-initiative system supporting integrated data exploration workflows, in: Proceedings of the Workshop on Human-In-the-Loop Data Analytics, ACM, 2018, p. 5.

[6] J.-D. Fekete, C. Plaisant, Interactive information visualization of a million items, in: Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on, IEEE, 2002, pp. 117–124.

[7] T. Siddiqui, A. Kim, J. Lee, K. Karahalios, A. Parameswaran, Effortless data exploration with zenvisage: an expressive and interactive visual analytics system, Proceedings of the VLDB Endowment 10 (4) (2016) 457–468.

[8] M. Khan, L. Xu, A. Nandi, J. M. Hellerstein, Data tweening: incremental visualization of data transforms, Proceedings of the VLDB Endowment 10 (6) (2017) 661–672.

[9] S. Amer-Yahia, B. Omidvar-Tehrani, J. Comba, V. Moreira, F. C. Zegarra, Exploration of user groups in vexus, ICDE demo.

[10] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, J. Heer, Vega-lite: A grammar of interactive graphics, IEEE transactions on visualization and computer graphics 23 (1) (2017) 341–350.

[11] M. Bhuiyan, S. Mukhopadhyay, M. A. Hasan, Interactive pattern mining on hidden data: a sampling-based solution, in: Proceedings of the 21st ACM international conference on Information and knowledge management, ACM, 2012, pp. 95–104.

[12] K. Dimitriadou, O. Papaemmanouil, Y. Diao, Aide: an active learning-based approach for interactive data exploration, IEEE Transactions on Knowledge and Data Engineering 28 (11) (2016) 2842–2856.

[13] N. Kamat, P. Jayachandran, K. Tunga, A. Nandi, Distributed and interactive cube exploration, in: Data Engineering (ICDE), 2014 IEEE 30th International Conference on, IEEE, 2014, pp. 472–483.

[14] B. Omidvar-Tehrani, S. Amer-Yahia, A. Termier, Interactive user group analysis, in: CIKM, ACM, 2015, pp. 403–412.

[15] L. van der Maaten, G. Hinton, Visualizing high-dimensional data using t-sne, Journal of Machine Learning Research 9: 25792605.

[16] E. Horvitz, Principles of mixed-initiative user interfaces, in: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, ACM, 1999, pp. 159–166.

[17] J. Heer, J. M. Hellerstein, Tutorial on data visualization and social data analysis, Proceedings of the VLDB Endowment 2 (2) (2009) 1656–1657.

[18] A. V. Pandey, A. Manivannan, O. Nov, M. Satterthwaite, E. Bertini, The persuasive power of data visualization, IEEE transactions on visualization and computer graphics 20 (12) (2014) 2211–2220.

[19] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, G. Melan con, Visual analytics: Definition, process, and challenges, in: Information Visualization – Human-Centered Issues and Perspectives, Springer, 2008, pp. 154–175.

[20] J. Lu, W. Chen, Y. Ma, J. Ke, Z. Li, F. Zhang, R. Maciejewski, Recent progress and trends in predictive visual analytics, Front. Comput. Sci. 11 (2) (2017) 192–207. doi:10.1007/s11704-016-6028-y.
URL https://doi.org/10.1007/s11704-016-6028-y

[21] M. H. Shimabukuro, E. F. Flores, F. Maria Cristina, de oliveira, and haim levkowitz,", in: Coordinated Views to Assist Exploration of Spatio-Temporal Data: A Case Study," 2nd International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV'04), 2004, pp. 107–117.

[22] Y. Chen, P. Xu, L. Ren, Sequence synopsis: Optimize visual summary of temporal event data, IEEE Transactions on Visualization and Computer Graphics.

[23] L. Wilkinson, The grammar of graphics, Springer Science & Business Media, 2006.

[24] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, D. A. Keim, Visual interaction with dimensionality reduction: A structured literature analysis, IEEE Transactions on Visualization and Computer Graphics 23 (1) (2017) 241–250. doi:10.1109/TVCG.2016.2598495.

[25] L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv e-prints.

[26] I. T. Jolliffe, Mathematical and statistical properties of population principal components, Principal Component Analysis (2002) 10–28.

[27] I. Borg, P. J. Groenen, Modern multidimensional scaling: Theory and applications, Springer Science & Business Media, 2005.

[28] P. Joia, F. Petronetto, L. Nonato, Uncovering representative groups in multidimensional projections, Computer Graphics Forum 34 (3) (2015) 281–290. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.12640, doi:10.1111/cgf.12640.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12640

[29] J. Wei, Z. Shen, N. Sundaresan, K.-L. Ma, Visual cluster exploration of web clickstream data, in: Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on, IEEE, 2012, pp. 3–12.

[30] J. Stasko, E. Zhang, Focus+ context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations, in: Information Visualization, 2000. InfoVis 2000. IEEE Symposium on, IEEE, 2000, pp. 57–65.

[31] P. Klemm, K. Lawonn, S. Glaßer, U. Niemann, K. Hegenscheid, H. Völzke, B. Preim, 3d regression heat map analysis of population study data, IEEE transactions on visualization and computer graphics 22 (1) (2016) 81–90.

[32] A. Makanju, S. Brooks, A. N. Zincir-Heywood, E. E. Milios, Logview: Visualizing event log clusters, in: Privacy, Security and Trust, 2008. PST'08. Sixth Annual Conference on, IEEE, 2008, pp. 99–108.

[33] J. Zhao, C. Collins, F. Chevalier, R. Balakrishnan, Interactive exploration of implicit and explicit relations in faceted datasets, IEEE Transactions on Visualization and Computer Graphics 19 (12) (2013) 2080–2089.

[34] X. Wang, Y. Liu, J. Lu, F. Xiong, G. Zhang, Trugrc: Trust-aware group recommendation with virtual coordinators, Future Generation Computer Systems 94 (2019) 224 – 236.

[35] B. Saket, H. Kim, E. T. Brown, A. Endert, Visualization by demonstration: an interaction paradigm for visual data exploration, IEEE transactions on visualization and computer graphics 23 (1) (2017) 331–340.

[36] D. Xin, X. Shen, Q. Mei, J. Han, Discovering interesting patterns through user's interactive feedback, in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2006, pp. 773–778.

[37] L. Boratto, S. Carta, G. Fenu, Discovery and representation of the preferences of automatically detected groups: Exploiting the link between group modeling and clustering, Future Generation Computer Systems 64 (2016) 165 – 174.

[38] G. Wang, X. Zhang, S. Tang, H. Zheng, B. Y. Zhao, Unsupervised clickstream clustering for user behavior analysis, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, 2016, pp. 225–236.

[39] M. Wattenberg, F. Vigas, I. Johnson, How to use t-sne effectively, Distill`doi: 10.23915/distill.00002`.
URL `http://distill.pub/2016/misread-tsne`

[40] B. Omidvar-Tehrani, D. Amer-Yahia, L. Lakshmanan, Cohort representation and exploration, in: Data Science and Advanced Analytics (DSAA), 2017 IEEE International Conference on, IEEE, 2018.

[41] S. Amer-Yahia, S. Kleisarchaki, N. K. Kolloju, L. V. S. Lakshmanan, R. H. Zamar, Exploring rated datasets with rating maps, in: Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017, 2017, pp. 1411–1419.

[42] C. Wilbert, Mental stimulation delays the start of memory decline, study shows, `http://www.webmd.com`.

[43] G. Charness, U. Gneezy, M. A. Kuhn, Experimental methods: Between-subject and within-subject design, Journal of Economic Behavior & Organization 81 (1) (2012) 1–8.

[44] J. Brooke, SUS: A quick and dirty usability scale, Taylor & Francis, 1996.

[45] B. Shneiderman, The eyes have it: A task by data type taxonomy for information visualizations, in: Visual Languages, 1996. Proceedings., IEEE Symposium on, IEEE, 1996, pp. 336–343.

[46] J. Sauro, A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices, CreateSpace Independent Publishing Platform, 2011.

[47] C. Ware, Information Visualization: Perception for Design, 3rd Edition, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2012.

[48] A. Graves, Techniques to reduce cluttering of rdf visualizations, Future Generation Computer Systems 53 (2015) 152 – 156.

[49] L. Jiang, P. Rahman, A. Nandi, Evaluating interactive data systems: Workloads, metrics, and guidelines, in: Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018, 2018, pp. 1637–1644.

[50] B. Omidvar-Tehrani, S. Amer-Yahia, Data pipelines for user group analytics, in: Proceedings of the 2019 International Conference on Management of

Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019., 2019, pp. 2048–2053.

[51] W. Mason, S. Suri, Conducting behavioral research on amazons mechanical turk, Behavior research methods 44.

[52] A. I. Chittilappilly, L. Chen, S. Amer-Yahia, A survey of general-purpose crowdsourcing techniques, IEEE Transactions on Knowledge and Data Engineering 28.

[53] A. Kittur, E. H. Chi, B. Suh, Crowdsourcing user studies with mechanical turk, in: SIGCHI, ACM, 2008, pp. 453–456.

[54] T. Blascheck, M. John, K. Kurzhals, S. Koch, T. Ertl, Va 2: a visual analytics approach for evaluating visual analytics applications, IEEE transactions on visualization and computer graphics 22 (1) (2016) 61–70.

[55] Oracle tpc benchmark, `http://www.tpc.org` (2017).

[56] O. Erling, A. Averbuch, J. Larriba-Pey, H. Chafi, A. Gubichev, A. Prat, M.-D. Pham, P. Boncz, The ldbc social network benchmark: Interactive workload, in: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, ACM, 2015, pp. 619–630.

[57] B. Omidvar-Tehrani, S. Amer-Yahia, User group analytics: Survey and research opportunities, TKDE.

[58] P. Eichmann, C. Binnig, T. Kraska, E. Zgraggen, Idebench: A benchmark for interactive data exploration, CoRR abs/1804.02593.

[59] T. Milo, A. Somech, Next-step suggestions for modern interactive data analysis platforms, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2018, pp. 576–585.