User Preferences for Hybrid Explanations

Pigi Kouki
UC Santa Cruz
pkouki@soe.ucsc.edu

James Schaffer
U.S. Army Research Laboratory
james.a.schaffer.ctr20@mail.mil

Jay Pujara
UC Santa Cruz
jay@cs.umd.edu

John O'Donovan
UC Santa Barbara
jod@cs.ucsb.edu

Lise Getoor
UC Santa Cruz
getoor@soe.ucsc.edu

ABSTRACT

Hybrid recommender systems combine several different sources of information to generate recommendations. These systems demonstrate improved accuracy compared to single-source recommendation strategies. However, hybrid recommendation strategies are inherently more complex than those that use a single source of information, and thus the process of explaining recommendations to users becomes more challenging. In this paper we describe a hybrid recommender system built on a probabilistic programming language, and discuss the benefits and challenges of explaining its recommendations to users. We perform a mixed model statistical analysis of user preferences for explanations in this system. Through an online user survey, we evaluate explanations for hybrid algorithms in a variety of text and visual, graph-based formats, that are either novel designs or derived from existing hybrid recommender systems.

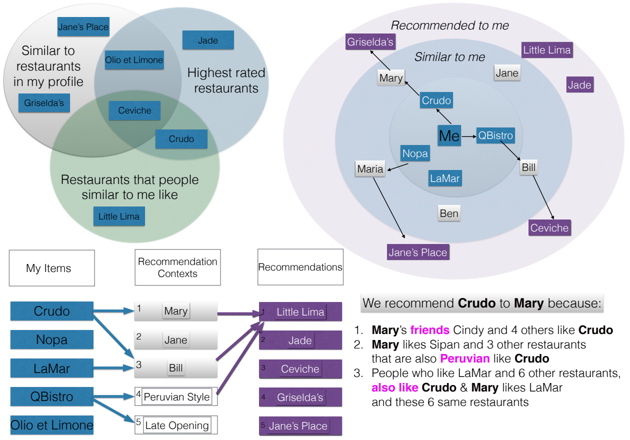
KEYWORDS

explanations, hybrid recommendations, hybrid explanations

1 INTRODUCTION

Successful recommender systems are integral to many applications ranging from movie recommendations to e-commerce. Effective recommendations must be both accurate and explainable [9, 10, 13, 16]. Hybrid recommender systems, which use a combination of signals such as social connections, item attributes, and user behavior, demonstrate improved recommendation accuracy [4]. However, compared to non-hybrid counterparts, hybrid models are more complex, and present many challenges when explaining recommendations to users. In this work, we study how to provide useful hybrid explanations that capture informative signals from a multitude of data sources without the burden of understanding the complex hybrid model.

Explanatory approaches for traditional, non-hybrid systems rely on a single explanation style (e.g., collaborative, content, knowledge, utility, or social explanation styles) [7]. Existing work has explored user preferences for these single-style explanations [2, 8].



We recommend **Crudo** to **Mary** because:

1. **Mary**'s **friends** Cindy and 4 others like **Crudo**
2. **Mary** likes Sipan and 3 other restaurants that are also **Peruvian** like **Crudo**
3. People who like LaMar and 6 other restaurants, **also like Crudo** & **Mary** likes LaMar and these 6 same restaurants

Figure 1: A subset of visualizations presented in our user study of hybrid explanations.

Visualization techniques for explaining recommendations include interfaces with concentric circles [11, 18], Venn diagrams [21], and pathways between columns [3], among many others. Hybrid explanations have been shown to be more effective than single-style explanations [20], but to the best of our knowledge, there is no study that adapts visualization techniques to hybrid recommenders or compares user preferences for hybrid explanations.

In this work, we identify several important dimensions for designing hybrid explanations. Using HyPER [14], a state-of-the-art recommender system, we develop a method for generating explanations from a hybrid model. We conduct a crowdsourced user study (N=200) to evaluate several different design approaches for hybrid explanations. This study answers several fundamental questions about designing hybrid explanations: 1) What visualization is best for hybrid explanations? 2) How should explanations be organized? 3) How much information should be in each explanation? 4) How detailed should each explanation be? Fig. 1 presents a sample of different visualizations that we generated in order to understand user preferences for hybrid explanations (for full details see [15]). This general evaluation strategy can be used to study user preferences for different recommendation domains, such as career sites, music services, and navigational routes.

Our contributions are: 1) a method for generating hybrid explanations from a hybrid recommender system based on a probabilistic programming language, 2) a list of ways explanations can be presented 3) a user evaluation of the different design approaches, and 4) a set of guidelines for designers of hybrid explanation interfaces.

RecSys'17, August 27–31, 2017, Como, Italy.

ISBN 978-1-4503-4652-8/17/08…$15.00.
DOI: http://dx.doi.org/10.1145/3109859.3109915

2 RELATED WORK

Several previous studies have considered explanations for recommender systems, demonstrating explanations can increase persuasiveness [8] and satisfaction [25, 26]. To better understand user preferences, research has compared explanation styles [2], combined different explanation styles [24], and considered dimensions such as personalization [25, 26], tags [30], ranking [17], and natural language presentations [5]. In addition to user preferences and design criteria for explanations, several GUIs and visualizations have been proposed for recommendations, including concentric circles (PeerChooser [11, 18]), clustermaps (TalkExplorer [29]), Venn diagrams (SetFusion [21]), and paths among columns (TasteWeights [3, 12]). Extending these ideas, we focus on *hybrid explanations* and study user preferences across designs and interfaces.

3 HYBRID EXPLANATIONS

The first challenge in understanding hybrid explanations is developing a system to translate a hybrid recommender model into a set of component signals and the utility of each component in the model. Here, we use HyPER [14] which makes use of probabilistic soft logic (PSL) [1], a generic probabilistic reasoning framework. We summarize the HyPER model, then describe our approach to converting the model's output to visualizable explanations. Our approach is compatible with any hybrid system with similar output.

The HyPER model is specified using a series of rules in first-order logic syntax defined by the modeler. The rules are used to define a probability distribution over recommendations. For example, to implement user-based collaborative filtering recommendations, the following rule is included in the PSL model:

$$w_{su} : \text{\textsc{SimilarUsers}}(u_1, u_2) \wedge \text{\textsc{Likes}}(u_1, i) \Rightarrow \text{\textsc{Likes}}(u_2, i)\ . \qquad (1)$$

Each rule is associated with a weight, learned from training data, capturing the importance of the rule to the model. In addition to the user-based collaborative filtering rule above, HyPER also includes item-based collaborative filtering rules with mean-centering priors, as well as rules for social-based recommendations using friendships, and content-based rules for using item attributes.

Next, a process known as *grounding* is used to combine the model with data and instantiate a set of propositions. For example, given a dataset with user ratings and a social network, user similarities, item similarities, social relationships, and item attributes are generated as evidence. Together, these ground rules are used by PSL to define a probabilistic graphical model which ranks unseen user-item pairs.

Running inference in HyPER generates recommendations captured by the $\text{\textsc{Likes}}$ predicate. After inference is complete, we select the top k items for each user. Then, for each of the top $\text{\textsc{Likes}}(u, i)$, we produce associated groundings used during the inference process. For example, in a restaurant recommendation setting, suppose that *Mary*'s top recommended restaurant is *Crudo* (i.e., the predicted value of the unobserved variable $\text{\textsc{Likes}}(Mary, Crudo)$ has the highest value among all other predicted values). While inferring the value of $\text{\textsc{Likes}}(Mary, Crudo)$, HyPER generated the following *ground* rules:

$$\begin{aligned}
\text{\textsc{Friends}}(Mary,\ Cindy) \wedge \text{\textsc{Likes}}(Cindy,\ Crudo) &\Rightarrow \text{\textsc{Likes}}(Mary,\ Crudo)\\
\text{\textsc{Peruvian}}(Limon,\ Crudo) \wedge \text{\textsc{Likes}}(Mary,\ Limon) &\Rightarrow \text{\textsc{Likes}}(Mary,\ Crudo)\\
\text{\textsc{SimilarUsers}}(Mary,\ John) \wedge \text{\textsc{Likes}}(John,\ Crudo) &\Rightarrow \text{\textsc{Likes}}(Mary,\ Crudo)\\
\text{\textsc{SimilarItems}}(Crudo,\ LaMar) \wedge \text{\textsc{Likes}}(Mary,\ LaMar) &\Rightarrow \text{\textsc{Likes}}(Mary,\ Crudo)
\end{aligned}$$

Explanation Style	We recommend *Crudo* because:
Social	1. Your friend Cindy likes Crudo
Content	2. You like Peruvian restaurants, like Crudo
User-based	3. Users with similar tastes as you like Crudo
Item-based	4. People who like LaMar, also like Crudo and you like LaMar
Item average rating	5. Crudo is highly rated
User average rating	6. You tend to give high ratings

Table 1: Example of an explanation for a restaurant (Crudo).

Since the groundings are different for each user-item prediction, the resulting explanations are inherently personalized. In Table 1, we present one way of visualizing these groundings using a parser that converts logical rules to natural language. In the next section, we discuss several design considerations and alternative presentations that can be implemented using the set of ground rules.

4 PRESENTATION OF EXPLANATIONS

Given hybrid explanations from HyPER, the next step is designing an interface to present these explanations to users. At a high level, the goal of any presentation style is to improve the user experience. We study the effect of different explanation presentation styles on user experience. To this end, we identify several dimensions for designing interfaces:

- **Presentation (Pres.)**: Natural language (Table 1), rule-based (equation 3), or graphical visualizations.
- **Weighting (Wgt)**: Whether or not explanation weights are displayed.
- **Grouping (Group)**: Whether or not explanations are grouped by style. Each rule can have many groundings in a dataset. For example, do users prefer to be shown the explanation *"Mary's friend Cindy likes Crudo; Mary's friend Josh likes Crudo"* or grouping explanations, *"Mary's friends Cindy and Josh like Crudo."*?
- **Information Density (Dens.)**: Amount of information shown.
- **Aggregation (Aggr.)**: Whether or not rules are aggregated in the grouping case. Do users prefer explanations of the type *"Mary's friends Cindy and 4 others like Crudo."* or *"Mary's friends Cindy, Josh, Rosie, George, Michael like Crudo."*?
- **Meta-explanations (Meta)**: Amount of high-level metadata in explanations (i.e., user similarity, item average rating, and user average rating).
- **Visualization (Visual)**: There are different ways of visualizing hybrid explanations, such as concentric circles [11, 18], Venn diagrams [21], and pathways among columns [3]. For Venn diagrams we used three intersections as suggested in [29]. Pathways among columns was tested with and without display of rules/reasoning.

In addition to the effect of different presentation styles on user experience, we also studied whether users have a specific preference over the *ranking* of the different explanation styles. For example, do users prefer to see social explanations before content-based ones?

5 EVALUATION

We constructed a set of 13 different treatments (Table 2) based on the explanation dimensions described in Section 4. Testing all possible subsets of dimensions was prohibitive, so we chose subsets that we judged as the most informative for explanation design. We

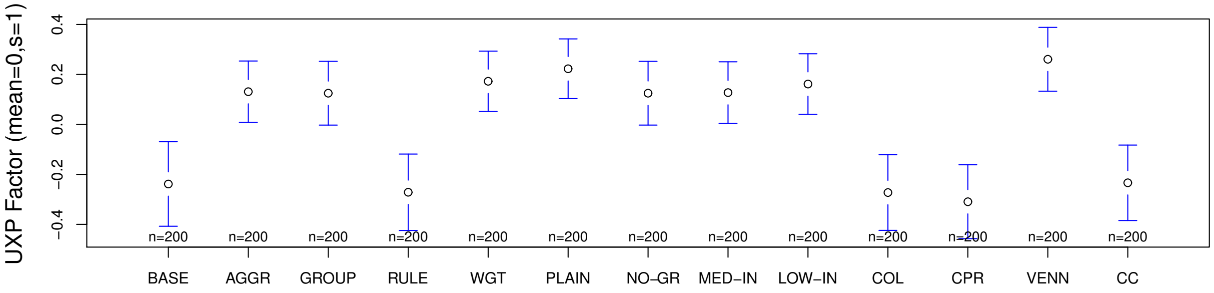


Figure 2: Mean UXP for each treatment. Errors bars are 95% confidence intervals. Treatment descriptions are given in Table 2.

Treat.	Pres.	Wgt	Group	Dens.	Aggr.	Meta
BASE	no exp.	no	no	NA	no	NA
AGGR	english	no	yes	low	yes	low
GROUP	english	no	yes	high	no	low
RULE	rule	no	no	low	no	high
WGT	english	yes	no	low	no	high
PLAIN	english	no	no	low	no	high
NO-GR	english	no	no	med	no	low
MED-IN	english	no	yes	med	no	low
LOW-IN	english	no	yes	low	no	low
		Visual Style				
COL	visual	COLumns + pathways				
CPR	visual	Columns + Pathways with Reasoning				
VENN	visual	VENN diagram				
CC	visual	Concentric Circles				

Table 2: Dimension values for each treatment for the different types of explanations tested.

evaluated explanations for a variety of textual and visual formats. We also included a baseline treatment (BASE) where we presented a recommendation item without any explanation.

We collected 200 samples of within-subjects participant data using Amazon's Mechanical Turk. The design used in our study [19] has been shown to minimize effects of satisficing (e.g., tab-click behavior) in crowdsourced studies. Overall, participants spent between 10 and 30 minutes for the study, 95% of participants were between 18 and 50 years of age, and 42% were male. The data was checked for satisficing users by checking input patterns and timings, however, none of the participants showed indications of violating the assumptions of the study. Each participant was rewarded with $0.5 as incentive.

5.1 Setup

Interface mockups were shown to participants in random order. All the mockups and details of the study can be found in [15]. We ran a synthetic experiment where all users were shown the exact same mockups that were manually generated. Each mockup presented a hybrid explanation for a random user called *"Mary"* for the restaurant *"Crudo"*. For each mockup, we elicited answers for a set of user experience questions, corresponding to understandability, system satisfaction, and perceived persuasiveness (Table 3).

Subjective metrics relating to recommender systems have shown to be strongly correlated (e.g., [22]), thus, confirmatory factor analysis was used to group the question items into a latent user experience variable to allow for simpler presentation of results and eliminate measurement error. A Cronbach's alpha [6] of 0.89 indicates good internal reliability of the constructed user experience (UXP) factor. Average variance extracted (AVE) was 0.64, indicating good convergent validity ($AVE > 0.5$).

Our experiment considered how participants' individual characteristics could affect user experience scores for each treatment. We asked pre-study questions related to visualization familiarity (VF, also shown in Table 3). Analysis showed co-variance between visualization familiarity and user experience was less than 0.5, which indicates good discriminant validity between the constructs.

Users were also given a task to rank different explanation styles according to their preference using a drag and drop interface. Specifically, we showed users the explanations from Table 1 (hiding the style) and asked them to rank these styles from the most to least important. We randomized the order different explanation styles were shown to avoid any order-related biases.

Factor	Question Item Description	R^2	Est.
VF	I am familiar with data visualization.	0.54	0.96
Cronbach: 0.85	I frequently tabulate data with computer software.	0.63	1.20
	I have graphed a lot of data in the past.	0.81	1.38
AVE: 0.60	I am an expert at data visualization.	0.57	1.21
UXP Cronbach: 0.87	*Understandability*: The recommendation process is clear to me.	0.73	0.86
AVE: 0.64	*Satisfaction*: I would enjoy using this system if it presented recommendations in this way.	0.68	0.82
	Persuasiveness: The recommendation is convincing.	0.77	0.88

Table 3: The latent VF and UXP factors built on participant responses to subjective questions.

5.2 Results

Figure 2 shows a plot of the mean user experience (factor loadings fixed to 1). To test for differences between the within-subjects treatments, we used structural equation modeling (SEM) [28], which can accommodate latent variables during significance testing, thus eliminating measurement error. We specified two factor models: the first with all within-subjects variables loaded onto a single factor (null hypothesis: no differences between treatments); the second with a factor specified for each of the 13 treatments (hypothesis: treatments cause a change in UXP). The model with a factor specified for each treatment achieved better fit. We used the Akaike Information Criterion (*AIC*) to estimate the quality of each model (a lower *AIC* indicates better comparative fit) and achieved

	BASE	AGGR	GROUP	RULE	WGT	PLAIN	NO-GR	MED-IN	LOW-IN	COL	CPR	VENN	CC
β	0.21	0.10	0.18	**0.41**	0.38	0.28	0.35	0.32	0.24	**0.57**	0.33	**0.39**	0.29
Err.	0.08	0.08	0.08	0.09	0.09	0.08	0.09	0.09	0.08	0.09	0.09	0.09	0.08
P	**	0.2	*	***	***	**	***	***	**	***	***	***	***

Table 4: Regressions coefficients (β) in a SEM that examine the relationship between visualization familiarity and observed user experience for each treatment. UXP/VF are latent variables with $\mu = 0, \sigma = 1$, effect sizes (β) on UXP are measured as SD from the mean as VF changes. Significance levels for this table: $^{***}\ p < .001$, $^{**}\ p < .01$, $^{*}\ p < .05$.

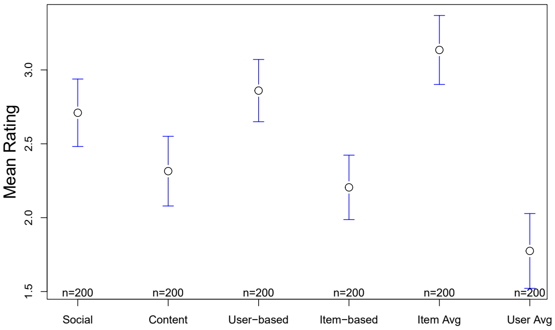


Figure 3: Means of participant ratings (1-6, with 6 being the highest preference) for the explanation style ranking task. Error bars are 95% confidence intervals.

$AIC = 25838$ for the single factor model vs. $AIC = 22908$ for the model with a factor for each treatment. This result indicates that there exist differences in UXP between treatments and thus the null hypothesis is rejected.

Next, we performed post-hoc tests between each treatment using a Raykov change model [23] (this mimics the popular Tukey test [27] while still allowing the use of latent variables). In this method, one factor (treatment) is used as a baseline and a slope is calculated between it and another factor. The post-hoc test showed that the interface using Venn diagrams (VENN) was significantly better than all other visual treatments and the baseline ($p < 0.001$ for BASE, COL,CPR, and CC). VENN also performed significantly better than AGGR, GROUP, NO-GR, MED-IN, and LOW-IN ($\forall\ p < 0.05$). The RULE treatment performed significantly worse than the explanations in plain English, as well as VENN ($\forall\ p < 0.001$). All English treatments performed significantly better than the baseline BASE ($\forall\ p < 0.001$). There was no significant difference between any of the English treatments ($\forall\ p > 0.10$), however, the mean for PLAIN was the highest. Consequently, weights, information density, aggregation, and grouping were also non-significant ($\forall\ p > 0.10$).

Results from the ranking task are shown in Figure 3. For our analysis we converted the ranking into rating data, i.e., the item listed first was given a rating of 6 and the item ranked last received rating 1. Users showed the strongest preference for item average rating explanations, followed by user-based and social explanations. A repeated measures ANOVA revealed differences in rating between the explanation styles ($p < 0.001$). A Tukey post-hoc test showed no statistical difference between social, user-based, and item average rating explanations ($\forall\ p > 0.10$). However, social, user-based, and item average rating were significantly better than item-based and user average rating ($\forall\ p < 0.05$). User-based and item average rating were significantly better than content explanations (both $p < 0.05$).

Finally, we conducted analysis on the relationship between visualization familiarity and user experience. A SEM was built with visualization familiarity regressed onto user experience measurements for each treatment (Table 4). Results indicate that visualization familiarity predicts increased user experience in all treatments, except AGGR. The highest increase in user experience is seen in COL and RULE. Model fit: $N = 200$ with 174 free parameters, $RMSEA = 0.064$ (CI : $[0.058, 0.069]$), $TLI = 0.89, CFI = 0.90$ over null baseline model, $\chi^2(772) = 1397$ (indicate acceptable fit, however, note that overall model fit is not an indicator of whether effects exist between variables).

6 DISCUSSION AND CONCLUSION

We have presented an evaluation of different visualization approaches using hybrid explanations. The results support prior findings [7] that explanations improve the user experience of recommender systems.

More specifically, Venn diagrams outperform all other visual interfaces and five of seven English interfaces, but are difficult to adapt to more than three sources. Natural language approaches were preferable to rule groundings from HyPER. Our experiments did not show a statistically significant difference across dimensions such as weights, information density, aggregation, or grouping in these explanations. This suggests that most plain English explanations may perform more or less the same in recommendation settings. Our results indicated that social, user-based, and item average rating explanations were the preferred explanation styles by users. Furthermore, we have established a reliable scale, as evidenced by Cronbach's α, for visualization familiarity which might be used to tailor explanation styles to individual users.

Additionally, we discovered that color-blind users ($N = 14$) rated the RULE interface higher in terms of UXP than the rest of the sampled population, with marginal significance ($\beta = 0.15, p = 0.051$). The mockup for this interface used a fairly intense violet color which may have been difficult to read for all but the colorblind users. While the colorblind sample was not large enough to change the results of the study, it highlights the need to accommodate these types of users when evaluating UXP for recommender systems.

The mockups that we showed to the users were manually produced and not generated by the HyPER system. As a result, the study was synthetic and did not support personalization. In future work we plan to analyze factors such as the quality of the recommendation and whether the user agrees and connects with the evidence. To this end, we plan to use HyPER's output, implement the best performing interfaces from our study (Venn, English PLAIN) and run a more comprehensive user case study in a lab setting to better understand effectiveness, efficiency, and user satisfaction of the explanatory system. One open question is determining effective methods for ranking explanations. We plan to compare several ranking strategies by designing interactive interfaces that support personalization.

REFERENCES

[1] S. Bach, M. Broecheler, B. Huang, and L. Getoor. 2017. Hinge-Loss Markov Random Fields and Probabilistic Soft Logic. *Journal of Machine Learning Research (JMLR)* (2017).

[2] M. Bilgic and R. Mooney. 2005. Explaining Recommendations: Satisfaction vs. Promotion. In *Proceedings of Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research at the 2005 International Conference on Intelligent User Interfaces*.

[3] S. Bostandjiev, J. O'Donovan, and T. Höllerer. 2012. TasteWeights: A Visual Interactive Hybrid Recommender System. In *Proceedings of the Sixth ACM Conference on Recommender Systems (RecSys '12)*. 35–42.

[4] R. Burke. 2007. Hybrid Web Recommender Systems. In *The Adaptive Web*. Springer.

[5] S. Chang, M. Harper, and L. Terveen. 2016. Crowd-Based Personalized Natural Language Explanations for Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. 175–182.

[6] L. Cronbach. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16, 3 (1951), 297–334.

[7] F. Gerhard and Z. Markus. 2011. A Taxonomy for Generating Explanations in Recommender Systems. *AI Magazine* 32, 3 (2011).

[8] J. Herlocker, J. Konstan, and J. Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. 241–250.

[9] D. Jannach and G. Adomavicius. 2016. Recommendations with a Purpose. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. 7–10.

[10] D. Jannach, P. Resnick, A. Tuzhilin, and M. Zanker. 2016. Recommender Systems — Beyond Matrix Completion. *Commun. ACM* 59, 11 (2016), 94–102.

[11] A. Kangasrääsiö, D. Glowacka, and S. Kaski. 2015. Improving Controllability and Predictability of Interactive Recommendation Interfaces for Exploratory Search. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. 247–251.

[12] B. Knijnenburg, S. Bostandjiev, J. O'Donovan, and A. Kobsa. 2012. Inspectability and Control in Social Recommenders. In *Proceedings of the Sixth ACM Conference on Recommender Systems (RecSys '12)*. 43–50.

[13] J. Konstan and J. Riedl. 2012. Recommender Systems: From Algorithms to User Experience. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 101–123.

[14] P. Kouki, S. Fakhraei, J. Foulds, M. Eirinaki, and L. Getoor. 2015. HyPER: A Flexible and Extensible Probabilistic Framework for Hybrid Recommender Systems. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15)*. 99–106.

[15] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, and L. Getoor. 2017. Supplemental Material: User Preferences for Hybrid Explanations. (2017). https://linqspub.soe.ucsc.edu/basilic/web/Publications/2017/kouki:recsys17:sup-mat/kouki-recsys17-sup-mat.pdf

[16] S. McNee, J. Riedl, and J. Konstan. 2006. Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)*. 1097–1101.

[17] K. Muhammad, A. Lawlor, and B. Smyth. 2016. On the Use of Opinionated Explanations to Rank and Justify Recommendations. In *Proceedings of the 29th International Florida Artificial Intelligence Research Society Conference (FLAIRS '16)*. 554–559.

[18] J. O'Donovan, B. Smyth, B. Gretarsson, S. Bostandjiev, and T. Höllerer. 2008. PeerChooser: Visual Interactive Recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. 1085–1088.

[19] D. Oppenheimer, T. Meyvis, and N. Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45, 4 (2009), 867–872.

[20] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos. 2012. A Generalized Taxonomy of Explanations Styles for Traditional and Social Recommender Systems. *Data Min. Knowl. Discov.* 24, 3 (2012), 555–583.

[21] D. Parra, P. Brusilovsky, and C. Trattner. 2014. See What You Want to See: Visual User-driven Approach for Hybrid Recommendation. In *Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI '14)*. 235–240.

[22] P. Pu, L. Chen, and R. Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the 5th ACM conference on Recommender systems (RecSys '11)*.

[23] T. Raykov. 1992. Structural models for studying correlates and predictors of change. *Australian Journal of Psychology* 44, 2 (1992), 101–112.

[24] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. 2009. MoviExplain: A Recommender System with Explanations. In *Proceedings of the Third ACM Conference on Recommender Systems (RecSys '09)*. 317–320.

[25] N. Tintarev and J. Masthoff. 2007. Effective Explanations of Recommendations: User-centered Design. In *Proceedings of the 1st ACM Conference on Recommender Systems (RecSys '07)*. 153–156.

[26] N. Tintarev and J. Masthoff. 2012. Evaluating the Effectiveness of Explanations for Recommender Systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 399–439.

[27] J. Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics* (1949), 99–114.

[28] J. Ullman and P. Bentler. 2003. *Structural equation modeling*. Wiley Online Library.

[29] K. Verbert, D. Parra, P. Brusilovsky, and E. Duval. 2013. Visualizing Recommendations to Support Exploration, Transparency and Controllability. In *Proceedings of the 18th International Conference on Intelligent User Interfaces (IUI '13)*. 351–362.

[30] J. Vig, S. Sen, and J. Riedl. 2009. Tagsplanations: Explaining Recommendations Using Tags. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI '09)*. 47–56.