Making Matches - Recommending the Right Personalities

Paul Schweiger
2103468
Multimedia Seminar II
Technische Universität Darmstadt

***Abstract*—Recommender Systems can be used to guide our limited time and attention towards meaningful actions, items, content or people. To tailor results to a user, Recommender Systems need to account for a user's preferences, skills and personality and incorporate these into a unified user model. These models need to be descriptive without overcomplicating the system. In social situations such as learning, recommendations become more complicated, as several users and their respective models need to be combined. Reciprocal recommendations can improve the quality of matchups by ensuring a good fit and benefit for both parties.**

A recent approach to implement a reciprocal peer recommendation platform for learning purposes by Potts et al. aims to show how these theories can be used in praxis. [1] The system enables students to find study-partners of matching skill for different learning topics by reciprocally recommending them to each other according to their respective preferences.

Based on this study, the intricacies of proper user modeling, reciprocal recommendations and researching new algorithms are highlighted. Common issues include unreliable self-reported user data, unevaluated algorithms and unfitting user models. Successful systems need to incorporate both explicit and implicit user feedback. Proper theoretical founding and preliminary considerations can solidify and improve reciprocal Recommender Systems.

I. Introduction

The modern world provides a plethora of opportunities, topics, people or products to consume, engage with or discuss, while people have a limited amount of attention and time to spend. As the offline world embraces online opportunities, this surplus of possible interactions is multiplied even further. In order to help us guide our attention and resources, basically every digital system tries to recommend meaningful content to users. Recommender Systems play a critical role in modern society and have transcended many different domains, pervading our lifes in advertising, e-learning, e-commerce, data analysis, online-dating, video game matchmaking, social networks, and others.

These highly unique topics make generalizing Recommender Systems difficult, leading to many different solutions for loosely related problems. [2] Combining some of these approaches, common techniques regarding tasks such as user modeling or social recommendation become prevalent. Employing Recommender Systems in social environments, helping users to engage with the right person at the right time, becomes a possibility that could shape interactions between people in lots of domains.

While social contacts prove highly important in many aspects of life, social and cooperative learning are also known to positively affect learning outcomes. [3], [4] Successful group learning efforts can enhance cognitive and intellectual performance, student's social and communicative skills and influence their overall satisfaction. [5], [6] Thus, social recommendation for learning is an important topic, helping to make mankind both more successful and happier at the same time.

To take advantage of social learning, research has diverted from the more traditional view of technology-enhanced learning, that focuses on improving individual learning experiences by recommending exercises, media resources or additional information at the right time depending on student's skills, preferences, needs and personality [7], [8]. Social learning research tries to connect learners to each other. Researchers have for example explored opportunities for students to receive immediate peer support via online requests [9] or tasked students to discuss their answers to exercises in an e-learning environment with peers, which led to improvements in both short- and long-term performance. [10]

The goal of this report is to emphasize the importance and common problems of proper user modeling and preference management for reciprocal social recommendations, especially in a learning context. Such Recommender Systems can not only be used to help a single learner solve problems or improve learning results, but to connect people, to build a community of learners and to enable students to engage in meaningful social learning opportunities. Along the way, relevant examples from other domains will be featured. After a quick introduction into the basics of user modeling and social recommendations in section II, section III will discuss a very recent paper by Potts et al. introducing a reciprocal Recommender System for learning environments. [1] Finally, section IV will discuss the paper's proposed prototype and highlight some of the intricacies of reciprocal recommendation and learning group formation.

II. Related Work: Recommender Systems

The goal of a Recommender System [RS] is to emphasize relevant pieces of information in a convoluted stream of data,

and to recommend a specific result to a specific user based on his or her history, preferences and situation. [11]
The following sections outline relevant findings and concepts regarding Recommender Systems in social environments and reciprocal peer recommendation. We will take a look at the basics of user modeling, social recommendations and the importance of reciprocality.

A. User Modeling

In order to successfully recommend items, an RS needs to understand it's user. Goals, circumstances and domain-specific aspects need to be considered. Thus, RS need to model their users to try and understand what items might be relevant. Recommender Systems need to account for an adequate operationalization of the relevant personality traits, domain-specific preferences and surrounding circumstances and combine these into a user model.
Different domains or approaches within the same domain require different user models. For example, information about people connected to the current user can be helpful to improve recommendation quality. Based on user friendships and shared interests, specific items that one person liked could be recommended to friends [12] or just other users with comparable interests via implicit user connections. [13] For example, a user's contacts and geographic history could be accessed to find users with matching profiles and to use their information to improve recommendations. [14]
For example in competitive multiplayer gaming, which is currently gaining in importance due to the increased interest in e-sports, the main goal is to create fair matches for two opposing teams. Usually, matchmaking in videogaming is concerned with bare player skills, but including the preferred style of playing, personality or character classes could lead to more balanced, fun games, making meaningful user models highly important.
Even within this single domain, a system could focus on self-reported preferences [15], ratings by other players [16], implicit interaction-derived data [17], [18] or a combination of technical and self-reported information to improve the overall gaming experience. [19]
A recent study by Wang et al. [20] looked at the enjoyment of multiplayer gaming sessions in League of Legends (LoL) based on player personality. They followed a subset of Sternberg's [21] problem-solving styles in order to categorize the playstyle of different users. Wang et al. automated the data collection process by using gameplay statistics. They assigned specific in-game actions to each problem solving style and determined a player's category based on his action profile. The results show a clear tendency of specific globally-active and risk-taking players to positively influence the overall game enjoyment, measured by the length of a game.
In theory, user models want to encompass as much personalized information as possible, without becoming so convoluted that item matching reverts back to being almost random. [22]
For instance, self-reported information like a user's basic information, preferences or even personality could be considered to improve the user model. [23] This personality data could then be used in many different applications, from recommending jobs to movies for a group to watch. [24], [25] Unfortunately, personalized data is hard to come by without having to ask the user directly, which can pose problems discussed in section IV.

B. Social Recommendation

In a world with an immense amount of possible social contacts, RS can help users to find other people to engage with, making recommendations relevant in social spaces. Lots of domain-specific social factors need to be considered in addition to the user model in applications in dating, learning, gaming, social networks or other domains.
In the educational sector, research has focused on building meaningful professional engagements. A system could, for example, find a supervisor who fits a student's needs in competence, personality and topic expertise. [26] Or one could try to help researchers find meaningful partners on academic conferences based on shared study-interests and personality. [27]
On a lower level of competency, studying with peers is considered an especially effective way to improve lots of different skills and build knowledge. [6] When engaging in higher education, many students move to a different town and thus lack a social environment. This makes finding a study group a huge initial challenge. Considering the many theories concerned with the effectiveness of learning group formation (heterogenous with different skill-levels and a minimum joint skill, diverse in terms of gender and ethnicity, ... [4], [28]), finding an actually helpful group seems to be impossible. This opens another field of study: Peer recommendation in learning, which will be the main focus of this report. [1], [22]
Contrary to the aforementioned topics, where oftentimes a specific match for one user had to be found, group learning has to be beneficial for everybody involved. This adds another layer of complexity to this kind of recommendation: Reciprocality.

C. Reciprocal Recommendation

In extreme cases, Recommender Systems will have to recommend users to each other, forming reciprocal recommendations: A user receives other users as recommendations and is himself an item recommended to others. A true reciprocal recommendation is found when two users are recommended to each other.
This does not have to be the main goal: Systems trying to recommend the best fit to each user without aiming for actual reciprocal recommendations could use simple scoring mechanisms. Each user receives scores describing their fit with other people. A certain amount of the highest scoring recommendations for each user will be returned. True reciprocal recommendations might happen as a byproduct of this process,

but are not enforced or pushed, as explained in [1] and in section III-B2.

When advantages for all participating users are aspired, true reciprocal recommendations become a necessity. Special modeling techniques have to be employed to boost the recommendation strength of reciprocal recommendations and make these more likely.

For example, Xia et al. successfully designed a Reciprocal RS for online dating, accessing self-reported user data and statistics of user's communication habits in the network. [29] To determine recommendation scores, they used a similarity-based approach between users, incorporating:

- the user's general profile information
- the user's willingness to communicate with others
- the user's attractiveness to others, derived from how many other people contacted him or her

As an earlier paper revealed, these implicit, behavioral details proved to be much more relevant to actually predict user interactions than self-reported preferences in dating partners. [30] Overall, the new system led to increased user satisfaction, since more informed and thus better matches could be made. [29]

III. Reciprocal Peer Recommendation for Learning Purposes

With the goal of providing opportunities for meaningful engagements between learners, benefiting mutual success, Potts et al. introduce a novel algorithm and platform for reciprocal peer recommendation in learning environments. [1] The scope of their study is to demonstrate the capabilities and explore the limitations of such a platform and algorithm on artificial data, before testing it under live conditions. Since the paper was published just recently before the writing of this report, further findings are not yet available. Large parts of their theoretical foundations have not undergone proper testing or focus on different domains, making the transfer of knowledge difficult and the results of this study highly interesting.

This chapter will cover the basics of their peer recommendation platform and evaluational findings on artificial data. We will then discuss some shortcomings of the paper at hand and delve deeper into learnings from other studies and topics that might benefit the overall performance of meaningful peer recommendation in the final section IV.

A. Recommendation in Personalised Peer Learning Environments [RiPPLE]

RiPPLE [”Recommendation in Personalised Peer Learning Environments”] was designed and developed as a web-based online learning recommendation system. *RiPPLE* is an adaptive, student-facing, open-source platform with the aim to enable students to engage with others in meaningful learning experiences. To enhance the learning experience, *RiPPLE* functions as a learning platform, helping students to co-create and find meaningful learning-content and to find peers to learn with. This analysis will focus on *RiPPLE* as a peer recommendation platform.

Based on user input, *RiPPLE* will calculate potential matchups for its users. Depending on

- the competency derived by a user's performance on learning content
- his or her available time slots
- the topics he or she would like to provide help, seek peer support or find a learning partner in and
- the user's preferences on the respective skills of a potential partner in these topics.

RiPPLE calculates a score for a matchup and will recommend a predefined amount of persons to each user. As *RiPPLE* currently recommends learning opportunities for the upcoming week, updates to user preferences or competencies are represented once per week.

An important aspect of the recommendation algorithm is it's compatibility function, calculating a one-directional score for each combination of potential study partners, u_1 and u_2. In the first step, the algorithm will check whether a potential matchup is viable following two hard constraints:

1) a shared time slot has to be available for both u_1 and u_2
2) the topic-specific joint competency must be greater than a predefined threshold τ. According to Blumenfeld [4], peer learning sessions will only become effective once the learners can share a minimum understanding of the topic.

For every pair of users satisfying these constraints, *RiPPLE* will calculate their respective one-directional scores. These represent how fitting u_2 is as a study partner for u_1 and vice-versa. (Since the users could have defined different preferences for their competency differences, scores don't need to be symmetric.) The scores take into account how good a matchup will be in terms of overall competency level, and how the other user matches the current user's preferences. These values will be calculated across all topics relevant for u_1 and u_2. A visual representation of the resulting score can be seen in figure 1.

These two one-directional scores could now be used to find the best partner for a specific user. To further recommend a matchup that is beneficial for both u_1 and u_2, the harmonic mean of both scores is considered as the ”reciprocal score” of u_1 and u_2, a value that is now symmetric. [31] The harmonic mean, contrary to the arithmetic mean, pays respect to differences between it's values, making a larger gap between values less desirable. Peer-combinations with approximately similar scores will receive better final values, making matchups that are beneficial to both participants more relevant.

In the last step, *RiPPLE* returns a predefined amount of matchups k with the best reciprocal values for each user. Although these reciprocal values are now symmetrical, the recommendations don't have to be: While from u_1's standpoint the matchup with u_2 and an (exemplary) reciprocal

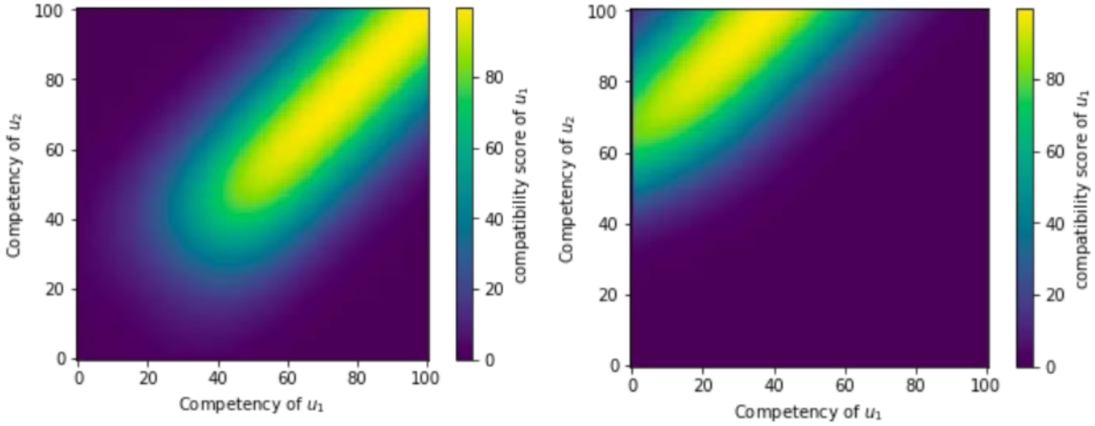

Fig. 1. The images show the areas of compatibility of a user u_2 as a function of u_1's competency. Lighter areas mean high compatibility scores in accordance to u_1's preferences. On the left u_1 is looking for a similarly skilled study partner, leading to the best fit along the $u_1 = u_2$ axis. The cutoff beneath a competency of 40 is due to the minimum joint competency threshold τ. On the right, we can see u_1 looking for peer support, i.e. a person with considerably higher knowledge, here about 60 points higher than u_1. Source: [1]

score of 30 could be the very best opportunity, u_2 could still have a true reciprocal matchup with u_3 and a value of 50.

For more information on *RiPPLE*, the algorithm and further clarification of different variables, please refer to [1].

B. Evaluation

In order to test *RiPPLE*'s applicability for actual use, Potts et al. designed an experimental setup in which *RiPPLE* would try to propose recommendations for randomly generated test data. Specific quality measures were designed according to [22] to assess different fields in which *RiPPLE* would have to show its capabilities. With satisfying results, *RiPPLE* would be able to be used under live conditions in the course of 2018.

For the experimental evaluation, random data had to be generated; diverse enough to highlight edge cases but within reasonable bounds.

To fully satisfy as a tool recommending students to one another, *RiPPLE* must be able to form meaningful and successful matches for as many users as possible in reasonable time. On the other hand, minor drawbacks in the defined metrics were considered to be tolerable in this step due to the experimental and randomly generated data and some further adjustments that could be made to compensate bad values.

As evaluation metrics for their experimental evaluation Potts et al. decided on four values that can further be used as general Quality Measures for reciprocal recommendation algorithms:

1) Scalability: With increasing enrollment numbers in higher education, *RiPPLE* will have to be suitable for large sets of learners. High runtime and costs for evaluating datasets with reasonable amounts of students means slower responses and a worse user experience. An optimal solution could provide immediate recommendations to any user, at any moment.

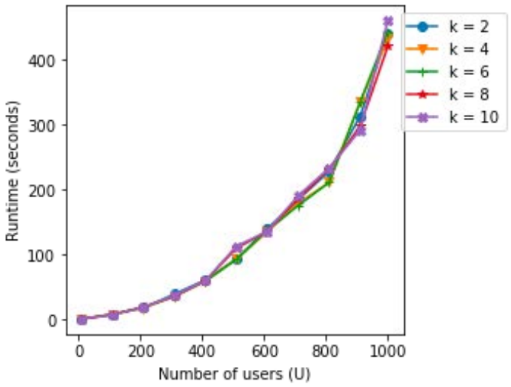

Fig. 2. Scalability: The algorithm's runtime depending on the number of users U and the amount of recommendations per user k. Note how k has almost no influence on the runtime, which grows exponentially with increasing U. Source: [1]

As can be seen in figure 2, the runtime of the algorithm increased in a quadratic fashion, as U, the total amount of users, increased: $O(n^2)$. The number of recommendations per user however did not significantly impact the runtime. (Although the paper states that it *did* in fact affect runtime, looking at the plots suggests that this might be a formatting error.)

Currently, *RiPPLE* calculates recommendations at the end of each week for the upcoming week, making the algorithm's runtime rather unimportant. In a 1000 user experiment, *RiPPLE* was able to provide recommendations for a single user in 0.045 seconds. However, further improvements are planned.

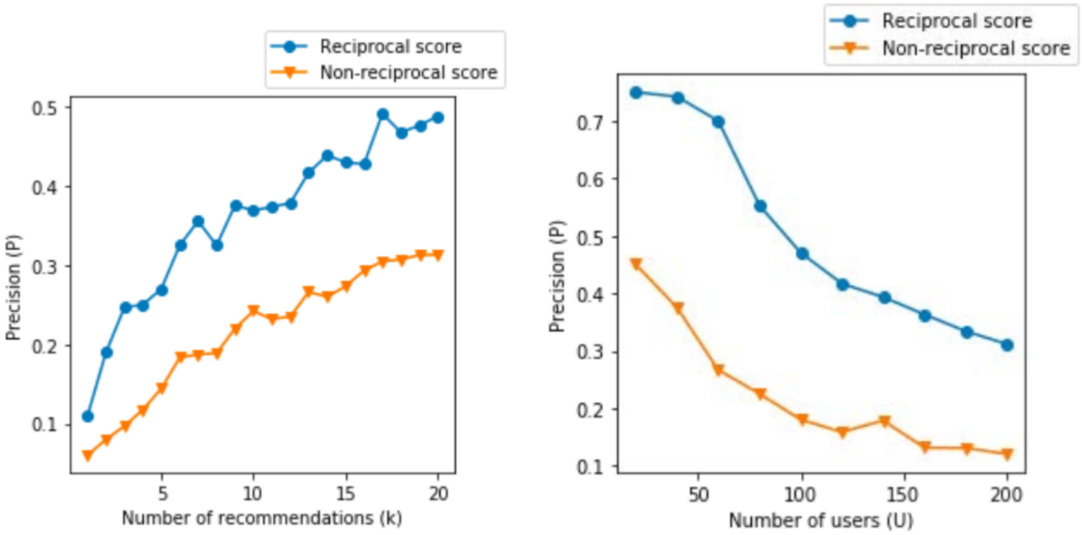

Fig. 3. Reciprocality: The precision (= the fraction of reciprocal recommendations out of the total recommendations averaged over all users) of the baseline non-reciprocal recommendations (orange) vs. of the reciprocal, averaged scores. (blue) Note how the reciprocal scores are always better. The left plot shows that increasing k also increases the precision, since more recommendations per user lead to a higher chance of reciprocal recommendations. On the right, increasing U with a fixed k reduces reciprocal precision, since there are more possible users to recommend. Source: [1]

2) Reciprocality: The best social recommendations are truly reciprocal: Users contacting a recommended user would also appear on this user's list of potential study partners. [31] Reciprocality was tested for both, the baseline non-reciprocal and the joint reciprocal harmonic mean scores. Whenever a user appears in the recommendations of a user on their own recommendation list that was built according to the respective score, the recommendation was considered to be reciprocal.

The precision for every user given the used score is calculated by dividing his reciprocal recommendations through k, the total amount of recommendations that user received. The system's total precision is defined as the average precision across all users. [31]

In all tested cases shown in figure 3, the reciprocal score had a higher precision than the baseline score. This is not surprising, since using the harmonic mean of both one-directional scores chooses reciprocal scores with medium values compared to non-reciprocal scores with a single high value. (As explained in section III-A)

3) Coverage: Recommending potential learning partners to one another should not abandon anyone. As such, coverage is a very important metric to consider. Since (almost) every user will receive recommendations, most users will be covered in one way or another. (The exception to this are users with completely incompatible time slots, role preferences (i.e. being the only person looking for an equally skilled study partner) or users who can't meet the minimum competency when coupled with their available potential partners.) A very good fit can only be ensured when each user is recommended to others, ideally forming a reciprocal recommendation, which is represented in metric III-B2. Coverage however is defined as the percentage of users that appear in other's recommendations at least once.

For a low amount of users and lots of recommendations per user, coverage was close to 0.9, meaning most users were recommended to others. As U increased or k decreased, the coverage sunk. However, more than 40% of users appeared in other's recommendations under all tested circumstances. Refer to figure 4 for a graphical overview.

4) Quality: The quality of a recommendation is not only based on its fit, but also on how good the resulting team could perform. According to Blumenfeld, learners should meet a minimum competency level in order to be an effective group, as specified by the employed minimum matchup threshold τ. [4] Quality is thus defined as the user's average joint competencies across their matched topics. The goal is to generate matches that are as capable as possible in their respective fields of study.

As figure 5 shows, it is apparent that the total amount of users did not affect the quality of matches. The minimum threshold for joint competency of a matchup however led to a better quality. Comparing this finding to figure 4 however, suggests that higher quality comes at the cost of less coverage. Especially when considering the slight improvements in quality score for larger increments in τ.

IV. Discussion

Reciprocal peer recommendation is a highly promising topic with lots of applications, but is hard to handle, as a closer examination of ”Reciprocal peer recommendation for Learning Purposes” and the implemented platform *RiPPLE*

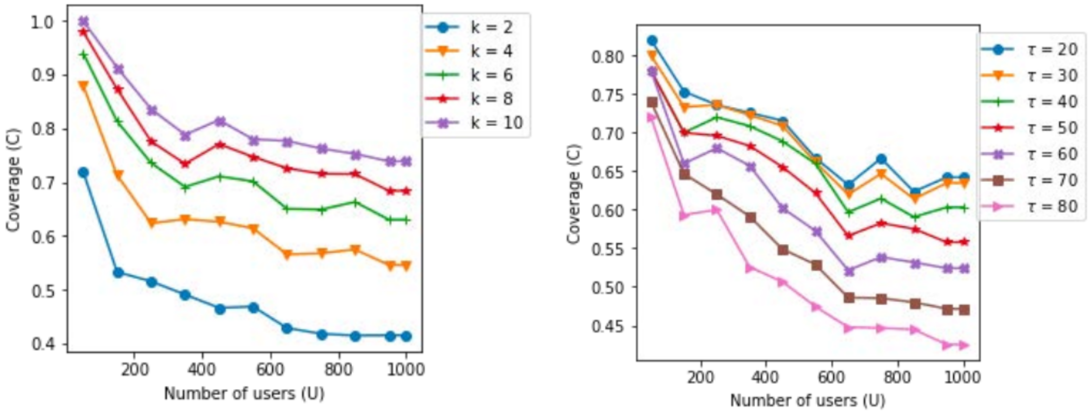

Fig. 4. Coverage: The percentage of users who appear in other's recommendations. With more users, coverage sinks (likelihood of hard to match users increases). Increasing the amount of received recommendations k (left) or lowering the minimum joint competency of matches (right) increases coverage. Mind the y-axis cutoff. Source: [1]

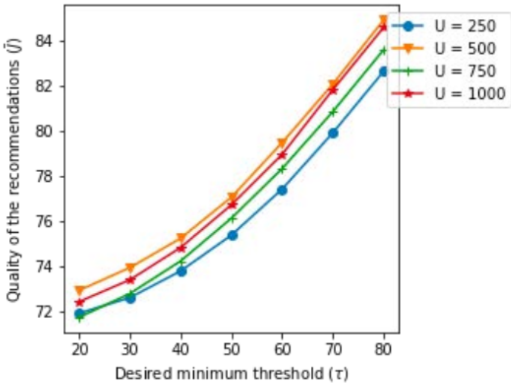

Fig. 5. Quality: The quality of matchups over different values for U and τ. Note how changing U does not affect matchup quality. Choosing a higher minimum joint competency threshold τ for successful matchups increases overall quality. Note the different scaling of the axes, overemphasizing the quality increases. Source: [1]

have shown. The system introduced by Potts et al. presents an approach to build a scalable, interactive and user-facing multi-purpose platform to enhance learning both in on- and offline circumstances.

The preliminary experimental evaluation of the platform's performance using artificial data sheds light on variable relationships, potential initial values for scientifically sparse concepts and the interplay of lots of factors. While the reported values present good metrics to measure the algorithm's performance and suitability for live data, they don't actually evaluate the algorithm, since no real data was considered and no target values have been set. Although the data is quantifiable, these findings should not be mistaken as quantitative results: instead of comparing the data to theory-driven goals and evaluating them for actual use, they are more or less providing an overview of how the algorithm works. In fact, there is currently no way to know if these results are ”good”. Some of the used measurements lack consistent data and research, and are not theoretically funded. For example, the question whether a coverage of little above 0.4 will be enough in practice, remains unanswered. Actual results from live usage are thus highly interesting and could provide insights in lots of different areas. The importance of user-centered on top of systemic evaluations has been discussed in several studies, specifically user satisfaction information might not be congruent with data-centric findings. [22], [32]

RiPPLE is concerned with theoretical criticisms regarding social learning RS research, and Potts et al. tried to pay respect to many problems. Summarizing many of these issues, a highly influential report by Olakanmi and Vassileva [22] highlights common shortcomings of peer recommendation studies and their proposed systems:

- Focus on improving learning instead of the goals of learning
- No information regarding the collection of user data
- Randomly assigning users in the first iteration and improving based on findings
- No consideration of scalability of the algorithm
- Inflexibility in working with fixed and limited constraints
- Too detailed user models leading to impossible matching
- Inflexibility in dealing with only partially known users
- Usage of self-reported user preference data
- Orphaned learners don't receive any value and have to be processed by hand
- No valid evaluation, just providing proofs of concept

RiPPLE was obviously built with these criticisms in mind, and tried to improve on the basis of earlier algorithms. For example, *RiPPLE* aims to provide actual benefits to learners, uses a detailed user model and assigns students as informed as possible. Scalability and many other factors were considered when reporting the platform's theoretical applicability in section III-B. *RiPPLE*'s flexibility in the definition of threshold values allows the tool to be customized to specific scenarios and needs.

However, many of the criticisms remain problematic until disproven: *RiPPLE*'s user model is light-weight but not evaluated and rather rigid, relying on the user's motivation to self-report data. The algorithm allows for learners to be orphaned due to different situations (as described below). Finally, as has already been stated, *RiPPLE* has not yet been evaluated in a real scenario.

Besides all this theoretical criticism, the algorithm itself does have some practical drawbacks as well. For one, *RiPPLE* calculates matchups across all topics. For example, two learners who would be a perfect match in one topic, but a bad match in another would be considered as a mediocre match. Topic-wise recommendation could further complicate the algorithm, but might lead to larger benefits for users.

Another way to improve the overall utility would be to consider slightly larger groups instead of matchups of two. Lots of theory about group composition emphasizes the importance of different skillsets and heterogeneity. [4], [22], [28] A system called DIANA, using a genetic framework to form small heterogeneous groups in study courses used many different student characteristics to match students. An evaluation using course grades proved that this could generally be more efficient than random or self-organized group formation. [33]

Another problem of *RiPPLE* are edge-cases in terms of competency preferences that would lead to overall low scores for matchups with other people, leading to learners who will receive suggestions with low scores, but won't appear in other's recommendations. While this does not necessarily lead to any consequences by itself, certain highly compatible users might be overwhelmed with meeting-requests from users outside their own recommendations. While they won't be able to meet every requesting user, these less compatible users might become abandoned.

Another drawback is the neglected human factor. Both user buy-in and competence in handling the tool and its demands might influence its use in practice. While this study's explicit goal was to only test the theory and future praxis tests are planned, this topic should still be discussed, a major shortcoming of the paper at hand.

A lack in user buy-in is something that should always be considered, especially in a student context. If a student didn't want to engage with strangers, was not motivated to study with partners or to adjust his or her schedule, all recommendations to and of that student would not accomplish anything. Meeting requests would be ignored, and opportunities for matchups would expire. Even negative user manipulation needs to be considered as a possibility, but is something that has to be dealt with in the live test.

While missed opportunities are a problem of the students themselves, rather than of the platform providing recommendations, the other human factor needs to be addressed directly by the tool:

As humans are unreliable, self-reported metrics always underly lots of variance and errors. A user's competency in a specific topic, his or her preferences, or the willingness to commit to a specific timeslot for learning might change daily, depending on mood, time of day, culture and lots of other factors. [34] [35] Other variables, like a user's preferred skill difference towards a learning partner, are especially hard to specify. How is a user supposed to know what his or her learning preferences are? How would he know which number refers to the desired difference in skill rating? From a psychological standpoint, this operationalization is bound to fail, if not controlled in an appropriate manner. [36]

This whole issue is not solely a problem of ”Reciprocal peer recommendation for Learning Purposes”. Generally speaking, choosing a user preference model that fits both the domain and the goal of an algorithm while still being able to perform in praxis poses a major problem in peer recommendation. [1], [22] In reciprocal recommendation, this issue becomes even more prevalent, since the user must be modeled as both, a recommended item and a user receiving recommendations at the same time.

An important factor is to consider both, the information needed to create meaningful matches according to a specific success criterion, and how to access this information. Usually, automatically collected data is preferred to relieve the users as much as possible. But not all data can automatically be collected. Internal information, like preferences, date availability, motivations or the highly important factor of personality need to be reported by the user - there is currently no way to easily access internal information of the user's mind.

Psychological research was founded to enable the measurement and quantification of these details. From psychophysical measurements to intricate operationalizations of complicated internal states, one easily accessible method stays predominant in social science: self-reported statements, from qualitative interviews to acceptance scales.

One such scale is the NEO-PI-R [37], arguably one of the most famous personality-measurement tools. A questionnaire with Likert-scaled [1] items on five axes representing the five dimensions of personality. [39], [40]

While the NEO-PI-R is widely used and considered as reliable and validated, it still relies on self- or peer-reported data and can, as such, not be considered a flawless tool. For example, many self-reporting tools suffer from a relevant problem called ”Faking Good”, the tendency to answer items in a way that is considered to be socially acceptable. When faking good, people manipulate their answers to cohere to social norms out of fear to be seen as a bad person. ”Faking Good” is known to influence the outcomes of some personality traits measured by the NEO-PI-R. [41] A similar effect can be observed when peer-ratings get influenced by peer sympathy. [42]

In learning environments, a phenomenon known as the ”Lake Wobegon Effect” influences student reports of their

[1] The commonly known items prompting users to reply to a statement on a scale from (usually) 1 to 5 (often combined with sentences like ”I agree” or ”I disagree”) are called ”Likert scales” after Rensis Likert. [38]

learning success: students tend to overstate their good performances, while failures will not be reported. This leads to an overestimation of student successes in surveys. [43]

As another example, the infamous ”Likert Scale” as introduced by Rensis Likert in 1932 as an ”attitude scale” [38] can safely be assumed to be one of the most used metrics in social research, while it’s optimal use still remains questionable. [34], [44]

Self-reported data is a largely controversial topic in psychological research: it is easy to acquire and enables researchers to access internal information, while these measurements can fluctuate following lots of different influences and are hard to validate. [34]–[36]

In summary, modeling users is complicated due to unreliable methods or participants (knowingly or subconsciously) manipulating their answers.

While Potts et al. obviously tried to choose a user-model that is limited to the necessary basics and tried to rely on as few ambiguous and self-reported pieces of information as possible, they *still* need a user’s ability in reporting information about him- or herself.

Other reciprocal recommendation approaches have to work around the same problems. As mentioned in section II-C, Xia et al. have shown that a behavior-emergent metric was more reliable in their use case of reciprocal online-dating recommendation. [30] Instead of focusing on user-reported data alone, they included implicit and pre-evaluated information derived from user interactions to measure attractiveness and willingness to communicate. [29] Wang et al. wanted to improve gaming matchmaking by employing problem-solving style information gathered from in-game statistics. They were able to deduce complicated, high-level cognitive problem-solving skills from simply evaluating implicit behavior, and used this information to improve player experiences. [20]

Instead of asking learners for their preferences regarding the competency difference towards their peer, Potts et al. could have decided on values founded in theory of optimal group composition. [28]

The contrary approach of data collection is to use implicit data derived from a user’s interaction with the running system. An important drawback to this method, mentioned by Olakanmi et al., needs to be considered: the cold start problem. [22] Relying on live data will lead to the first recommendations to be made without any underlying data, at random. Only after some burn-in, actual data can be used to achieve better results, which leads to much less initial acceptance for the new tool.

Another drawback of implicit data is finding a correct operationalization of relevant variables. As shown in section II-A, systems in a straightforward domain such as gaming matchmaking can choose many completely different approaches to measure good experiences and fair matches for players. An approach valid in one situation might be completely ill fitting after changing the game, genre, game mode, users or goal of the RS. Finding viable measurements that are reliable and valid is one of the major concerns of psychological experimental design. Implicit data might be more reliable, but is hard to design.

But what kind of data is better for any given Recommender System? The discussion about implicit versus explicit data acquisition and user feedback is still ongoing. As has been established in the former paragraphs, while explicit data is easy to collect but often biased by subjective inabilities to properly report it, implicit data is more objective but harder to acquire, especially for some specific bits of data.

In another recent study, fittingly named ”Explicit or Implicit Feedback? [...]” [45], Zhao et al. compared different recommendation algorithms in different performance criteria. They collected implicit (user interactions with the recommended items) and explicit (user feedback) information, but used different subsets for training their RSs. Their main finding was that machine-learning algorithms optimized on implicit action prediction led to higher engagement results from users. Systems learning with explicit user feedback however, led to more user satisfaction in comparison. Unsurprisingly, the system maximized what it was told to maximize. However, users were more satisfied with a system that was not concerned with engagement rates, but with overall quality, even if this might incorporate less interactions. The authors conclude that this might have contributed to late research focusing on implicit feedback and neglecting the actual benefit for a user while focusing on the higher engagement rates, which include both negative and positive interactions. While still lots of studies focus primarily on systemic data instead of actually testing their prototypes with users [32], an interaction between the endorsement of high engagement rates and low evaluation expenses emerges. However, as Zhao et al.’s work confirms, explicit user feedback is still considered to be highly important to actually monitor user satisfaction and system performance. [2], [32]

The last and probably most important finding of Zhao et al.’s work was that a combined data approach worked even better: implicit and explicit data taken together can often outperform regular systems. As other research confirmed, checking for user’s feelings about a recommendation and validating it by watching implicit data delivers more reliable data than any singular approach could. [13], [32], [45]

V. Conclusion

As no truly satisfying approach to choose data and models for effective group formation in a new and untouched cohort of peers has yet been found, Potts et al. struggle to compile a convincing user model themselves. They incorporated explicit and implicit user data into their system, but could not avoid all of the pitfalls of data collection.

Taking all of this into consideration, their proposed peer learner recommendation algorithm does have its flaws but could still benefit learners in praxis. Reciprocally matching students according to their needs and preferences is likely to help at least some students achieve better grades and form new social connections.

The applications of reciprocal peer recommendation are manifold and generally beneficial in a highly social and connected modern world. Bringing different people together and providing opportunities to engage with one another yields many advantages and teaches meaningful skills. However, human factors need to be incorporated in the algorithms to account for the dualistic role of users as both recommended items and recipients of recommendations. Many experiments with explicitly and implicitly collected data have uncovered a vast array of issues that need to be navigated. Especially due to the many problems associated with measuring humans, algorithms are still unreliable when handling people in reciprocal situations. However, skillful design can circumvent some of these problems and create systems that are far from perfect, but nonetheless beneficial to their users. Time will tell if *RiPPLE's* streamlined user model and up-front evaluation of technical measures have made it a platform fit for practical use. Until then, further research is required.

References

[1] B. Potts, H. Khosravi, C. Reidsema, A. Bakharia, M. Belonogoff, and M. Fleming, "Reciprocal peer recommendation for learning purposes," in *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 2018.

[2] J. Buder and C. Schwind, "Learning with personalized recommender systems: A psychological view," *Computers in Human Behavior*, vol. 28, no. 1, pp. 207 – 216, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0747563211001956

[3] S. T. Bossert, D. C. Dwyer, B. Rowan, and G. V. Lee, "The instructional management role of the principal," *Educational administration quarterly*, vol. 18, no. 3, pp. 34–64, 1982.

[4] P. C. Blumenfeld, R. W. Marx, E. Soloway, and J. Krajcik, "Learning with peers: From small group cooperation to collaborative communities," *Educational researcher*, vol. 25, no. 8, pp. 37–39, 1996.

[5] C.-M. Zhao and G. D. Kuh, "Adding value: Learning communities and student engagement," *Research in higher education*, vol. 45, no. 2, pp. 115–138, 2004.

[6] J. Maxwell, "Learning together: Peer tutoring in higher education." *Journal of Physical Therapy Education*, vol. 22, no. 3, p. 97, 2008.

[7] H. Drachsler, K. Verbert, O. C. Santos, and N. Manouselis, "Panorama of recommender systems to support learning," in *Recommender systems handbook*. Springer, 2015, pp. 421–451.

[8] M. Erdt, A. Fernandez, and C. Rensing, "Evaluating recommender systems for technology enhanced learning: a quantitative survey," *IEEE Transactions on Learning Technologies*, vol. 8, no. 4, pp. 326–344, 2015.

[9] J. Greer, G. McCalla, J. Cooke, J. Collins, V. Kumar, A. Bishop, and J. Vassileva, "The intelligent helpdesk: Supporting peer-help in a university course," in *International Conference on Intelligent Tutoring Systems*. Springer, 1998, pp. 494–503.

[10] C. A. Reidsema, L. Kavanagh, E. Ollila, S. Otte, and J. E. McCredden, "Exploring the quality and effectiveness of online, focused peer discussions using the moocchat tool," in *27th Australasian Association for Engineering Education Conference*. AAEE, 2016.

[11] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook," in *Recommender systems handbook*. Springer, 2011, pp. 1–35.

[12] H. Feng and X. Qian, "Recommendation via user's personality and social contextual," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2013, pp. 1521–1524.

[13] C. C. Hsu, M. Y. Yeh, and S. d. Lin, "A general framework for implicit and explicit social recommendation," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2018.

[14] L. Ramaswamy, P. Deepak, R. Polavarapu, K. Gunasekera, D. Garg, K. Visweswariah, and S. Kalyanaraman, "Caesar: A context-aware, social recommender system for low-end mobile devices," in *Mobile Data Management: Systems, Services and Middleware, 2009. MDM'09. Tenth International Conference on*. IEEE, 2009, pp. 338–347.

[15] J. Riegelsberger, S. Counts, S. D. Farnham, and B. C. Philips, "Personality matters: Incorporating detailed user attributes and preferences into the matchmaking process," in *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*. IEEE, 2007, pp. 87–87.

[16] W. I. Patrick, S. D. Lamb, M. Bortnik, and J. P. Hansen, "System and method for providing feedback on game players and enhancing social matchmaking," 2011.

[17] M. Suznjevic, M. Matijasevic, and J. Konfic, "Application context based algorithm for player skill evaluation in moba games," in *Network and Systems Support for Games (NetGames), 2015 International Workshop on*. IEEE, 2015, pp. 1–6.

[18] O. Delalleau, E. Contal, E. Thibodeau-Laufer, R. C. Ferrari, Y. Bengio, and F. Zhang, "Beyond skill rating: Advanced matchmaking in ghost recon online," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 4, no. 3, pp. 167–177, 2012.

[19] S. D. Farnham, B. C. Phillips, S. L. Tiernan, K. Steury, W. B. Fulton, and J. Riegelsberger, "Method for online game matchmaking using play style information," 2009.

[20] H. Wang, H.-T. Yang, and C.-T. Sun, "Thinking style and team competition game performance and enjoyment," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 7, no. 3, pp. 243–254, 2015.

[21] R. J. Sternberg, *Thinking styles*. Cambridge University Press, 1999.

[22] O. A. Olakanmi and J. Vassileva, "Group matching for peer mentorship in small groups," in *CYTED-RITOS International Workshop on Groupware*. Springer, 2017, pp. 65–80.

[23] M. A. S. Nunes and R. Hu, "Personality-based recommender systems: an overview," in *Proceedings of the sixth ACM conference on Recommender systems*. ACM, 2012, pp. 5–6.

[24] P. T. Costa Jr, R. R. McCrae, and G. G. Kay, "Persons, places, and personality: Career assessment using the revised neo personality inventory," *Journal of Career Assessment*, vol. 3, no. 2, pp. 123–139, 1995.

[25] J. A. Recio-Garcia, G. Jimenez-Diaz, A. A. Sanchez-Ruiz, and B. Diaz-Agudo, "Personality aware recommendations to groups," in *Proceedings of the third ACM conference on Recommender systems*. ACM, 2009, pp. 325–328.

[26] M. Zhang, J. Sun, J. Ma, T. Wu, and Z. Liu, "A personality matching-aided approach for supervisor recommendation (research-in-progress)," in *System Sciences (HICSS), 2016 49th Hawaii International Conference on*. IEEE, 2016, pp. 678–687.

[27] N. Y. Asabere, A. Acakpovi, and M. Michael, "Improving socially-aware recommendation accuracy through personality," *IEEE Transactions on Affective Computing*, 2017.

[28] S. Manske, T. Hecking, U. Hoppe, I.-A. Chounta, and S. Werneburg, "Using differences to make a difference: A study in heterogeneity of learning groups," in *11th International Conference on Computer Supported Collaborative Learning (CSCL 2015)*, 2015.

[29] P. Xia, B. Liu, Y. Sun, and C. Chen, "Reciprocal recommendation system for online dating," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. ACM, 2015, pp. 234–241.

[30] P. Xia, K. Tu, B. Ribeiro, H. Jiang, X. Wang, C. Chen, B. Liu, and D. Towsley, "Characterization of user online dating behavior and preference on a large online dating site," in *Social Network Analysis-Community Detection and Evolution*. Springer, 2014, pp. 193–217.

[31] S. Prabhakar, G. Spanakis, and O. Zaiane, "Reciprocal recommender system for learners in massive open online courses (moocs)," in *International Conference on Web-Based Learning*. Springer, 2017, pp. 157–167.

[32] S. Fazeli, H. Drachsler, M. Bitter-Rijpkema, F. Brouns, W. van der Vegt, and P. B. Sloep, "User-centric evaluation of recommender systems in social learning platforms: Accuracy is just the tip of the iceberg," *IEEE Transactions on Learning Technologies*, pp. 1–1, 2017.

[33] J. Moreno, D. A. Ovalle, and R. M. Vicari, "A genetic algorithm approach for group formation in collaborative learning considering multiple student characteristics," *Computers & Education*, vol. 58, no. 1, pp. 560–569, 2012.

[34] J. W. Lee, P. S. Jones, Y. Mineyama, and X. E. Zhang, "Cultural differences in responses to a likert scale," *Research in nursing & health*, vol. 25, no. 4, pp. 295–306, 2002.

[35] J. Sorensen, "Measuring emotions in a consumer decision-making contextapproaching or avoiding," *Aalborg University, Denmark*, 2008.

[36] R. M. Gonyea, "Self-reported data in institutional research: Review and recommendations," *New directions for institutional research*, vol. 2005, no. 127, pp. 73–89, 2005.

[37] F. Ostendorf and A. Angleitner, *NEO-Persönlichkeitsinventar nach Costa und McCrae, Revidierte Fassung (NEO-PI-R)*. Hogrefe Göttingen, 2004.

[38] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932.

[39] R. R. McCrae and P. T. Costa, "Validation of the five-factor model of personality across instruments and observers." *Journal of personality and social psychology*, vol. 52, no. 1, p. 81, 1987.

[40] L. R. Goldberg, "An alternative" description of personality": the big-five factor structure." *Journal of personality and social psychology*, vol. 59, no. 6, p. 1216, 1990.

[41] B. Griffin, B. Hesketh, and D. Grayson, "Applicants faking good: evidence of item bias in the neo pi-r," *Personality and Individual Differences*, vol. 36, no. 7, pp. 1545–1558, 2004.

[42] D. Leising, J. Erbs, and U. Fritz, "The letter of recommendation effect in informant ratings of personality." *Journal of Personality and Social Psychology*, vol. 98, no. 4, p. 668, 2010.

[43] N. L. Maxwell and J. S. Lopus, "The lake wobegon effect in student self-reported data," *The American Economic Review*, vol. 84, no. 2, pp. 201–205, 1994.

[44] L. Chang, "A psychometric evaluation of 4-point and 6-point likert-type scales in relation to reliability and validity," *Applied psychological measurement*, vol. 18, no. 3, pp. 205–215, 1994.

[45] Q. Zhao, F. M. Harper, G. Adomavicius, and J. A. Konstan, "Explicit or implicit feedback? engagement or satisfaction?" 2018.