

Ratings, Reputations, and Prediction Markets for Peer Patent Examination

Paul Resnick

Initial draft: 10/31/06

Current draft: 12/3/06

This document is prepared in response to the questions about Rating, Ranking, and Reputations for the Peer2Patent project, posted at http://www.communitypatent.org/project_docs/ratings_and_rankings/index.html. Thanks to Rahul Sami for helping me think through the application of prediction markets, and to Beth Noveck for pushing me to make the connections with the actual design problems of the P2Patent project clearer.

This proposal is primarily concerned with incentives for *Contributors*¹ to provide information, and ways to aggregate and present that information to inform other Contributors and Examiners².

The range of potential inputs from Contributors includes:

- Prior art references
- Textual comments
- Tags
- Ratings or rankings³
- Predictions/bets

The charge for this workshop is to suggest a design, drawing on one or more of the elements above, to accomplish two informational goals:

1. Focus attention and labor of the community of Contributors on the patents and claims where it is most needed.
2. Generate a ranked list of prior art, possibly limited to 10 items.

There is also one incentive-oriented goal:

3. Encourage Contributors to contribute productively.

Overall, the approach I will suggest has the following features:

- For allocating Contributor attention, rely heavily on the tagging system and on linking comments to claims and prior arts references.

¹ Contributor is one of the role names used in the project documentation. Contributors are registered with the site and provide commentary and suggestions of prior art relevant to applications. See http://www.communitypatent.org/use_cases/actors/index.html for a listing of other actor types.

² Examiners are Patent Examiners, employed by the US PTO.

³ I will use the word “ranking” to refer to an ordering of items. A “rating” will be used more broadly for any explicit expression of user opinion about an item. If every item gets a distinct numeric rating, then the items can be sorted in order to determine a ranking.

- For generating a ranked list of prior art, aggregate predictions made by Contributors, with Contributors' predictions weighted more heavily if they have more credibility. Credibility builds based on quality of past predictions.
- For encouraging productive contribution, display metrics of participation, including the credibility scores.
- Notably absent from these suggestions is any subjective quality rating by Contributors of claims, comments, prior art, or Contributors. I think it would be unclear to Contributors what the rating criteria were supposed to be, and hence such ratings would be difficult to interpret.

I consider each of the three goals, in order. Each section begins with the questions posed to us by Beth Noveck, in italics, followed by my response. An Appendix documents in more detail the suggestion about how to generate credibility scores and how to aggregate predictions using the credibility scores.

1. Focus Attention Among Claims

Ranking the claims of a patent application to identify the most relevant/representative ones. PURPOSE/GOAL: focus community attention and labor where most needed

What kind of system should be used to rank? Is a simple up or down vote good enough? It would be difficult to have the users put many claims into a particular order. What about a system of tokens to identify most important claims (e.g. 3 tokens, each one placed on the most important claim)

I don't think that a rating or ranking system is the best way to allocate Contributor attention among patent claims. One big problem with using ratings or rankings for this is that Contributors will have different scales in their heads—some will rate on the quality of the claim (i.e., do they think it will stand up to Examiner scrutiny); others will rate on the perceived importance of prior art to invalidate the claim (i.e., a very broad claim that the Contributor is worried will have a chilling effect on innovation).

Moreover, I think that a system of tagging and linking comments and prior art to specific claims will allow more informative displays than any aggregation of subjective ratings would do. Imagine a display of the claims for a single patent that showed, along with each claim, some summary statistics, including⁴:

- A tag cloud of tags (weighted by frequency) that have been applied to the claim;
- A count of how many comments have been linked to the tag, and how many of those comments the current Contributor has not yet read, with links to view them;
- A count of how many prior art references are linked to the claim, with a link to view them;
- The current community prediction about the probability that the Examiner will disallow or require an amendment of the claim.

⁴ To avoid biasing Examiners, some or all of these statistics might be hidden from Examiners' views of the community input related to a patent.

Community social norms could arise around tags to use (e.g., “need prior art citation”), without pre-judging all the tags that Contributors might use to guide each other’s attention allocation. The count of commenting activity related to the comments would also give an indicator of where other people were focusing their attention.

The system design already envisions prior art references being linked to patent claims.⁵ This same idea merely needs to be extended to allow for linking of comments to claims. After entering a comment, a user would be given a list of claims that the system thinks the comment applies to. The user could manually change that list, but frequently would just click to accept the system’s suggestion. The system would suggest claims that a comment should link to in two ways. First, if a comment is a direct reply to another comment, any claims the original comment was linked to would be checked off as relevant to the new comment. Second, a computer program would scan the text of the comment to identify any textual references to claims. I think this would require a relatively simple parser to match text of the form “claim <n>” or “claim[s] <n>, <n>...” or “claim[s] <n.>-<m>”.⁶

The community prediction about the probability that the Examiner will disallow or require amendment of a claim is generated from a prediction market. Participants make predictions assigning a probability x% that the Examiner will disallow or require amendment of this claim. Details of how the individual predictions are aggregated to form a consensus prediction can be found in an appendix.

2. Generate a Ranked List of Prior Art

*Ranking by peer reviewers of **prior art** submitted by the community in response to a patent application. PURPOSE/GOAL: create manageable and searchable output for patent examiner*

What kind of system should be used to rank? What scale should we use? How should those ranking be displayed to the reviewers? Should we randomize the order in which the prior art is presented so that distribution of ratings is fair? In other words, should we require/request that people rank a submission before submitting prior art or a comment of their own? Should we allow changing of votes once submitted? How do we avoid last minute gaming of the system? Is IP-address tracking enough to solve this problem? What will go into the ranking (e.g. will number of votes be relevant)? Will there be a minimum number of votes necessary for something to be ranked in the list? What about tied submissions?

⁵ See p. 7 of http://www.communitypatent.org/project_docs/files/p2patent_mockup_Nov2006.pdf

⁶ The MovieLens system has included a similar text parser to identify references to movies within free text of a discussion forum, with user verification. Check out the site MovieLens.org, or see the paper, “Insert movie reference here: a system to bridge conversation and item-oriented web sites”, available through the ACM digital library at <http://doi.acm.org/10.1145/1124772.1124914>.

Since the purpose is to highlight the prior art most likely to be relevant, what we really want is the community consensus about what will be most useful to the patent examiner. While some kind of rating or ranking system might generate a reasonable approximation, a prediction market is designed to produce exactly that, so it seems like the right tool for the task. Subsequent action by the examiner, citing or not citing the item, will reveal whether the community was correct, and thus determine payoffs to bettors in the market.

Each piece of suggested prior art has its own prediction market. At any point in time, the market for a piece of prior art has a number between 0% and 100% which indicates a probability that the Examiner will cite it in the final application or rejection. Any Contributor can submit her own prediction, between 0 and 100, which causes the communal probability to be updated. The maximum amount that a single Contributor can move the communal consensus is limited by the Contributor's credibility score, which is determined from the accuracy of her previous predictions. Details of the limits and credibility score calculations can be found in the Appendix.

At the time when the market closes, each piece of prior art has a consensus probability score. The prior art can be ranked in descending order of those scores. If a cutoff of only the top 10 items is desirable, then only the items with the ten highest percentage predictions would be shown to the Examiner. So as not to bias the examiner, the actual scores need not be revealed to the Examiner.

The major threat to the system is the introduction of spurious prior art that is not relevant to the application, thus crowding out prior art that is relevant. The algorithm for limiting the effect of a Contributor's prediction makes it more difficult to carry this out. Several Contributors will all have to give low predictions for relevant prior art, or high predictions for irrelevant prior art, in order to knock a relevant item out of the top 10 that will be presented to the Examiner. Moreover, the attacker will have to build up its credibility before even being able to participate in the attack.

This proposal is, however, vulnerable to an attacker who creates many identities (call them sock puppets) and strategically makes predictions with them so as to build up their credibility. The problem is more acute if the system assigns "provisional credibility" based on agreement with other Contributors before the Examiners make rulings. The Appendix discusses this in more detail. I believe this is not a weakness of the particular design presented here but rather an inherent problem for any rating or ranking system or prediction market, if it allows anonymous participation and gives any weight to opinions before getting objective evidence from Examiners about the raters' credibility.

3. Encourage Contributors to contribute productively

Rating of community participant activity. PURPOSE/GOAL: to encourage the right kind of participation.

Possible modes of rating/ranking:

- a) automatically based on submission of prior art or comments;*
- b) rating by other participants of each other,*

*c) rating of members of reviewers by the patent examiner,
d) rating of members automatically based on the citation of their prior art and commentary on the file wrapper of the patent*

With regard to 3) (rating participation)

What scale/system shall we use for rating the quality of participation in the system?

3)(a): (automatic)

Should the system award points for submissions and/or for comments? Do submissions and comments have to be approved by the moderator before points are awarded? Does the eventual ranking of the submission ascribe more points to the submitter? How does the rating change when they submit 10, 50, 500 prior art items?

3)(b): (ranking of participants by other participants)

Do we want a rating systems by participants of each other based on subjective criteria? Based on objective criteria? Would this create a greater incentive to participation?

3)(c): (rating by patent examiner)

Do we want patent examiners to rate (or give a thumbs up) to people whose submissions/comments were particularly useful? Of course, there are legal/political questions to resolve with regard to whether this is possible but, technically, is it desirable?

3)(d): (rating of members automatically based on file wrapper)

We imagine that if a piece of prior art is cited by the examiner on the final application or rejection, extra points will accrue to the reviewer who submitted the prior art and who submitted supporting comments. This will require labeling comments as being in support of or in opposition to a submission.

I think various metrics about people should be presented, in several ways:

- Summary metrics for contributors should be shown alongside each of their comments or prior art suggestions; it should be possible to click through to a user profile;
- The user profile should include more detailed metrics and links to the Contributor's history of activities;
- Leaderboards should show leaders on various metrics, to encourage a spirit of competition. Some of the metrics should be based on short time windows, in order to give everyone a sense that they have a chance to make it onto the leaderboard.

As to the content of the metrics, I think they should be based on a pass-through of feedback about a Contributor's specific actions, rather than direct assessments of the Contributor. Since all the activity of Contributors is visible in the site, it makes more sense to evaluate that activity than to have general evaluations of people. (Contrast this with eBay, for example, where buyer and seller interact offline—even there, the site evolved from users rating other users to users ratings specific transactions.) Thus, I don't like 3)(b) or 3)(c). I do, however, think the following might be useful metrics:

- Number of comments, prior art citations, patents commented on, tags entered, etc
- Average communal prediction for prior art first suggested by this user
- Credibility score for predictions about prior art and about claim rejections [this is 3)(d) in the questions posed to us]
- To the extent that Examiners can be asked to issue assessments of particular comments, some metric based on those assessments would also be useful.

The question posed to us seems to suggest generating a single composite metric, with some points for making comments, some points for entering prior art, etc. I'm not convinced that a composite metric will be especially motivating to Contributors or informative to others.

Appendix A: Prediction Markets for Prior Art and Claim Validity

The prediction markets for prior art and claim validity work similarly. In either case, the output of the market is a community assignment of a probability to the event that the patent examiner will take a particular action (cite a piece of prior art or disallow a claim, depending on the market). Different community members will have made different predictions about how the examiner will act, putting at stake some of their reputation points, so effectively the predictions are bets. The examiner's action will determine the point payoffs of those predictions. For simplicity of exposition, the remainder of the appendix focuses on the prediction market for prior art.

At each instant in time, there is a current probability assigned to the outcome that the item of prior art will be cited. Initially, the probability is 0. The first person to suggest the item will give a prediction that will move the communal consensus.

The bettor can move the communal assessment about a particular piece of prior art, in either direction. The quadratic scoring rule, detailed below, determines a score associated with each communal assessment. The payoff to a bettor is the score for the probability they chose minus the score for the probability that they moved it from. Thus, bettors are rewarded for moving the prediction in the direction of the outcome that actually occurs. This mechanism is known as a market scoring rule [Robin Hanson, "Combinatorial Information Market Design", *Information Systems Frontiers*, 5(1), pp. 1387-3326. 2003].

Quadratic Scoring Rule

Let $P = \text{pr}(\text{item will be cited})$

Let $x = 1$ if the item is cited; $x = 0$ otherwise.

$S(P, x) = 2P - [P*P + (1-P)*(1-P)]$ if $x=1$

$S(P, x) = 2(1-P) - [P*P + (1-P)*(1-P)]$ if $x=0$

Market Quadratic Scoring Rule

Let P_1 be the communal assessment before the bettor acts.

Let P_2 be the probability she chooses.

Credibility points gained = $S(P_2, x) - S(P_1, x)$

For example, suppose the current communal consensus is that the probability of citing the item is $P_1=1/4$. The bettor thinks it should be $P_2=1/2$. If the item is actually cited, the payoff will be

$S(P_2, 1) - S(P_1, 1) =$

$2(1/2) - [1/4 + 1/4] - (2(1/4) - [1/16 + 9/16]) =$

$1/2 - (-1/8) =$

.625

If, on the other hand, the item is not cited, the payoff will be

$$\begin{aligned} S(P2, 0) - S(P1, 0) &= \\ (2(1/2) - [1/4 + 1/4]) - (2(3/4) - [1/16 + 9/16]) &= \\ 1/2 - 7/8 &= \\ -.375 \end{aligned}$$

Before we consider a few modifications to make it better fit the current application, consider the basic incentive properties. If the bettor moves the communal assessment in the direction of the eventual outcome, she gains points; otherwise she loses points. If a bettor thinks that the communal probability prediction is wrong, she has an expected gain of points from correcting it. Moreover, to maximize her expected payoff, she will move it to exactly her true beliefs about the probability the claim will be denied. This property of the quadratic scoring rule is well known in the literature and makes it a *proper scoring rule*. [See e.g., Axiomatic Characterization of the Quadratic Scoring Rule, Reinhard Selten, *Experimental Economics* 1 (1), pp.43-61].

Any linear multiple of the scoring rule has the same incentive properties (payoffs are maximized with honest reports of true beliefs); the constant may be chosen as part of the system design to determine how many good predictions a Contributor needs to make to build up a high credibility score.

If there is a community consensus around a particular probability, then any move from that probability creates an arbitrage opportunity for someone else—by correcting the move, they can gain points in expectation by predicting the original community consensus. Thus, the current public probability assessment should be a reasonable aggregation of the views of the peer reviewers, as anything else leaves an opportunity for someone to make a bet with positive expected value.

The scoring rule can equivalently be viewed equivalently as a betting mechanism where larger bets get lower payoff odds. But I think that it will be much more natural for Contributors to simply state their beliefs about what the Examiner will do than to place bets about the outcome with the optimal bet size depending on the Contributor's beliefs.

There are a few improvements we can make, to make the system simpler for users to understand, more resistant to manipulation, and able to provide tentative scores for bettors during the time after a bet is placed and before the patent examiner acts. We detail these below.

Resistance to Manipulation

The system can cap the amount that any single user can move the communal prediction for an item of prior art. There may be a generic cap applying to everyone, and there may be personalized caps determined by an individual's current point total. A person who has accumulated a reputation as a good predictor, as indicated by a high point total, will be able to make a larger change to the communal prediction. The larger the change a bettor

makes to the communal prediction, the more points she can lose (or gain). If we set the bettor's current point total as the maximum number of points she can lose, that determines a maximum change to the communal prediction that she can make.

Caps can limit the influence of a single user, but will not help if people are able to create many accounts. Each account might manipulate the prediction only by a little bit, but together they might manipulate by a lot. This kind of attack is known in the computer science literature as a sock puppet or sybil attack [Cheng and Friedman, Sybilproof reputation mechanisms, Proceeding of the 2005 ACM SIGCOMM workshop on Economics of peer-to-peer systems, pp. 128-132.]. We propose two counter-measures to this kind of attack.

The first is to require a registration process that limits the creation of new accounts. One individual should not be allowed to create more than one account, or at least not too many accounts. We leave the details of the registration validation process to others to work out—it might be that limits based on IP addresses would be sufficient, or it might be that people would be required to provide evidence of their true names, even though the names were not revealed to participants in the system.

The second is to start new users with no credibility in the system, so that they cannot influence the communal assessment until they have proved themselves to be credible predictors. That is, a newcomer will be able to make bets and these bets pay off *as if* the bettor was actually moving the communal prediction. A somewhat weaker version of this is to provide Contributors who have gone through some external vetting process with a small initial credibility score, so that they can start to have some immediate influence.

More generally, we now allow a divergence between the effect of a bet on the communal assessment and the effect of a bet on the bettor. The number of points that a person has accumulated so far will limit the amount that they can affect the communal prediction, but for the purposes of building their own credibility score the system will allow them to place larger bets *as if* they were having a larger affect on the communal prediction.

Note that this is one way that the total supply of points in the system can increase—several bettors can be paid as if they made a large change in the communal prediction, because those bettors without established reputations will not affect the communal prediction by much, thus leaving room for another bettor to again make a large bet. (A bet can only be large if it makes a large move in the communal probability assessment, from P1 to P2).

Timing Issues: Provisional Payoffs

Typically, a long time will elapse between the placement of a bet and the patent examiner ruling on an application. Since the action of the patent examiner determines the outcome of the bet, that means a bettor's point total is unknown. But the point totals are used to determine how much influence a user can have on communal probability assessments; initially a bettor will have little or no influence. Thus, initially the system will not make communal assessments that differ from the initial assessment. This may be appropriate as

a conservative approach (after all, we don't know how good the peer patent process will be at identifying items that should be cited), but it also means that the system provides little guidance to patent examiners until it has been proven, which may be problematic. Moreover, credibility points will have very little motivating effect if it takes years before payoffs are known.

One potential solution is to offer provisional payoffs on bets, based on the current communal probability assessment. The provisional payoff on a bet is the expected value of the bet given the current assessment of the probability that the item will be cited. That is, if the current assessment is $\text{pr}(\text{item cited}) = p$, then the payoff on any bet would be $p*[S(P2,1)-S(P1,1)] + (1-p)[S(P2,0)-S(P1,0)]$. As communal assessments change, provisional payoffs can change. As in a stock market bubble, the communal consensus may be wrong and thus provisionally reward the people who initiated that incorrect consensus. Unlike a stock market bubble, however, the correct payoffs will eventually be determined, when the patent examiner makes a ruling.

In order to jump-start the ability of communal assessments to form, some participants would need to start with credibility scores greater than 0. This could be done by waiting for patent examiners to rule on some patents, or, more likely, by simply giving some points to some of the initial participants.

At each point in time, a user's provisional reputation score is the sum of the payoffs for the bets that patent examiners have ruled on plus the provisional payoffs for those not yet ruled on. The provisional reputation scores can be used to determine how much influence users can have with their bets on other items. That is, at any point in time, the sequence of bets made by the bettors can be rerun, with the amount of influence that any bettor can have being capped by the bettor's provisional reputation score.

As described in the section on manipulation, those caps will affect the impact of bets on the communal prediction, but will not affect the nature of the bets for payoff purposes. If a user places a bet to move the prediction from current level of P1 to P2, she will be paid off based on that bet. Subsequently, because of changes to provisional reputations, when the market for the current claim is rerun, the communal consensus prior to this user's bet may no longer be P1. And the impact of her bet on the communal consensus may not be to move it to P2. But she will still be paid off as if she had moved the bet from P1 to P2.

We note that this system of provisional reputations may induce new opportunities for collusive manipulation. For example, user A places a bet and user B places a bet that creates a positive provisional payoff for user A. User A then follows B in betting on some other item to create a positive provisional payoff for B. We speculate that this danger of collusive reputation inflation is an inherent danger induced by any system that rewards based on consensus with other raters rather than rulings by the patent examiner. Further analysis is needed to determine whether any such reputation inflation can be eliminated once the patent examiner does rule. Even if it is eventually removed, a bad actor can temporarily have significant influence; if it can then create additional fake identities as needed, the system will be vulnerable.

There are some details still to be worked out and some formal analyses of incentive properties that would be worth doing. I postpone those until it becomes clear that a prediction market of some kind is potentially desirable for this project.