# Evaluating Play Trace (Dis)similarity Metrics

**Joseph C. Osborn** and **Ben Samuel** and **Joshua Allen McCoy** and **Michael Mateas**
University of California, Santa Cruz
{jcosborn,bsamuel,mccoyjo,michaelm}@soe.ucsc.edu

## Abstract

Play trace dissimilarity metrics compare two plays of a game and describe how different they are from each other. But how can we evaluate these metrics? Are some more accurate than others for a particular game, or in general? If so, why? Is the appropriate metric for a given game determined by certain characteristics of the game's design? This work provides an experimental methodology for validating play trace dissimilarity metrics for conformance to game designers' perception of play trace difference. We apply this method to a game-independent metric called Gamalyzer and compare it against three baselines which are representative of commonly used techniques in game analytics. We find that Gamalyzer—with an appropriate input encoding—is more accurate than the baseline metrics for the specific game under consideration, but simpler metrics based on event counting perform nearly as well for this game.

## Introduction

It is difficult for a game designer to predict what will happen when their game is in players' hands. During the early phases of design, it is feasible for a designer to directly observe players and make changes accordingly; but as the number of players increases, this ad hoc analysis cannot scale.

Accordingly, the academy and the games industry (motivated by design concerns as well as business requirements such as profitability and user retention) have invested substantial effort in gathering and analyzing game play data (el Nasr, Drachen, and Canossa 2013). A natural artifact to examine is the *play trace*, a sequence of player actions corresponding to one play of the game.

Many design questions for popular genres such as first-person shooters concern the game's spaces (and are thus amenable to spatial heatmaps) (Kim et al. 2008). Unfortunately, there are many genres (e.g. puzzle games) where spatial superposition of game states is unhelpful, and there are many design questions which are difficult to answer just by looking at color densities. To avoid committing to particular features of a given game's states, we propose that the *difference between play traces* is an effective general-purpose measurement which can be used to help answer a variety of design and player-modeling questions.

- "Do players pursue diverse strategies?"
- "Are winning traces similar to each other?"
- "What are the outlier plays of this game?"
- "Is it possible to win without being at all similar to this canonical trace?"

There are many metrics for computing play trace similarity, both game-specific and game-independent. For example, n-gram counts of actions could be compared or the terminal states of those traces could be compared. These comparisons abstract over play traces: the former considers unordered sets of counts and the latter merges the whole sequence into a single state. One recently developed metric, Gamalyzer (Osborn and Mateas 2014), applies a variant of edit distance to compare play traces directly, without abstracting over time. While we can imagine several similarity metrics, we should use the most correct one: that which agrees the best with the designer's own perception of differences between play traces.

In this paper we work directly with the designers of Prom Week (McCoy et al. 2013) to validate Gamalyzer's ability to answer questions of interest. To this end we develop instruments to validate the claim that any given play trace dissimilarity metric is capturing the same kinds of differences as a human designer. We can imagine that the better a metric agrees with human appraisals of difference, the more useful it will be in answering questions like the ones above.

A valid underlying distance metric is necessary but not sufficient to validate an operational definition of something like strategic diversity, overall uniqueness, or outlier detection. On the other hand, if we happen to have a sound operational definition that uses one distance metric, we can gauge the quality of some other metric by swapping it in and seeing if the accuracy of this definition improves or degrades.

This paper has two primary hypotheses. We first hypothesize that because Gamalyzer (and operational definitions based on it) considers whole play traces it will be more accurate than general-purpose measurements like n-gram counting that lose the temporal context of events. Our second hypothesis is that Gamalyzer will dominate approaches based on comparing ad hoc features of game states or play traces,

because these tend to ignore symmetries in a game's design (alternative paths which reach the same destination) and because it is difficult to select the correct state features even with expert knowledge. Finding support for these two hypotheses with respect to a given game would imply that Gamalyzer is a valid play trace dissimilarity metric for that game. This would provide some evidence for the informal claims of generality in the original Gamalyzer paper (Osborn and Mateas 2014).

## Related Work

### Comparing Play Traces

Representative examples of play trace visualization and analysis include histograms of game event counts and BioWare's overlaying of player actions (including meta-game actions like asking other players for help) onto a game's map (el Nasr, Drachen, and Canossa 2013). These visualizations, software, and analyses are generally invented as needed to help answer a particular design question for a particular game.

Playtracer (Andersen et al. 2010) is one example of a design support tool which directly compares play traces (specifically, sequences of game states). It is a visualization that neither maps state onto a game's navigational space nor is genre-, game-, or query-specific. Unfortunately, Playtracer does not include a general-purpose game state dissimilarity metric, and incorporating Playtracer into a design process requires that designers both identify relevant state features and define a state distance metric using those features; these problems can be difficult even for experts.

It is important to note that we are not evaluating the usability of visualizations or user interfaces in this paper; only that the conclusions drawn by these automated processes are *correct* with respect to the designers' expert knowlege. This is in some sense both easier and harder than developing a usable interface, but it is often taken for granted.

### Evaluating Distance Metrics

How should a distance metric be evaluated? Some distance metrics are completely hermetic: string edit distance, for example, is exactly the least expensive set of changes to turn one string into another. This purely syntactic measure would be unaware of the similarity between "green" and "viridian."

Often (and in the particular case of play traces) we want to measure the difference between two objects of interest in terms of some *semantic* qualities (such as player experience or strategy). Other domains have this property as well: while there are many ways for computers to compare two images for similarity, if we want to present the results to humans we should pick a metric that agrees well with human perception.

A particularly thorough investigation of image comparison metrics comes from Rogowitz, Frese, Smith, Bouman, and Kalin (1998). In this study, two psychophysical experiments were performed: one in which humans arranged 97 images on a table such that the distance between two images stood for the dissimilarity of those two images; and one in which humans repeatedly selected which of a subset of the 97 images was most similar to a reference image from the remainder of the images. Both of these experiments produced similarity matrices (the first complete and the second sparse) which were reduced to a low-dimensional space in pursuit of the most important perceptual features of images.

In designing our experiments, we supposed that one feature that play trace comparison has in common with image comparison is that most distances, on a 0 to 1 scale, are likely to be close to 1: so different as to be effectively incomparable. The sparse distance matrix in the image perception experiments matched the complete distance matrix quite closely—despite the extremely different experimental setup—because most of the entries in the matrix are 1.

In our case, we are not trying to derive the features that humans use to discern play trace differences; instead, we are trying to validate that the distances found by various metrics conform to the distances found by humans. The experiments conducted by Russel and Sinha comparing the L1 and L2 distances for image dissimilarity (2011) are a closer match for the aims of this work. Here, the authors also controlled for semantic content so that human ratings purely concerned the visual properties (rather than the subjects) of the image. Semantic content is the whole point of play trace analysis, so our experimental design borrows from both studies.

We would like to note that there are other meaningful kinds of play trace differences besides the designer perception of player strategy, including for example differences in play style. Evaluating metrics' suitability for those purposes is outside the scope of this paper, but the techniques we show here should be broadly applicable.

## About Prom Week

Gameplay in Prom Week revolves around the social lives of 18 characters at a high school in the week before their senior dance. Each scenario (or *story*) in Prom Week centers around the social goals of one character. For example, the goals of Chloe's story (an introductory level) are to help her make peace with a notorious bully and to start dating the boy she has always loved from afar.

The player works toward goals by selecting characters to engage in *social exchanges*—patterns of social behavior that change the social state. Each exchange is categorized in one of several *social intents*, which are high-level social goals. The social exchanges available (and the likelihood of their success) are determined by the *volitions* of the characters. In our Chloe example, players might want her to engage in the social exchange "Ask Out" with her crush and "Make Peace" with her bully. A recent journal article gives an in-depth description of social exchanges and the AI system that drives Prom Week (McCoy et al. 2014).

Each Prom Week play trace file contains a list of all social exchanges played in that particular trace; from this, the social state at any given timestep can be reconstructed. Since its initial release on February 14, 2012, over a hundred thousand play trace files have been generated, making Prom Week a good candidate for developing novel forms of evaluation (Samuel et al. 2014). Links to play the game for free can be found at `promweekgame.com`.

## About Gamalyzer

Gamalyzer is a variant of the constraint continuous edit distance applied to game play traces (Osborn and Mateas 2014). Briefly, it finds the cheapest way to turn one play trace into another using only matches, insertions, and deletions. The cost of matching a single game event (or input) to a different game event is defined by a recursion on the name and parameters of that event. This syntactic difference is interpreted as a semantic difference, because the encoding of play traces as sequences of parameterized game actions depends critically on design knowledge to determine the types and names of events, their parameters, et cetera.

Gamalyzer assumes that play is goal-directed, that substantial differences in length indicate substantial semantic differences, that inputs arrive at roughly the same rate in every trace, and that events which are far apart in time are incomparable. This last assumption is the *constraint* in constraint continuous edit distance: a parameter called the *warp window* ($\omega$) prevents match operations for pairs of inputs of each trace which are too far apart.

So far, Gamalyzer's outputs have been rationalized on an ad hoc basis, and it seems to produce valid output from synthetic data. This paper lays out the first tests of its accuracy as a distance metric with respect to a *designer's* conception of play trace difference using actual data. Moreover, Prom Week is distinct enough from the platformers and puzzle games used in Gamalyzer's debut that the metric's performance herein should provide evidence for the claim that Gamalyzer is game-independent.

Gamalyzer encodings of game inputs consist of two main parts: a determinant and a value. If two inputs have different determinants, they are incomparable (their change cost is infinity); if their determinants match, then their values are compared recursively, with some parts of the value contributing more significantly to change cost than others. The determinants and value are sometimes called the *parameters* of an event, one of which (generally the first parameter of the determinant) is the *event type* or *name*.

For this paper, we evaluate Gamalyzer with two encodings of Prom Week play traces. Each move in Prom Week has the player select an initiating character, a social exchange, and a target character. In the first encoding ($\texttt{glz}_{\texttt{ie>t}}$), every input has the same determinant (`move`); the social exchange's initiating character (*i*) has the same relevance as the combination of social exchange (*e*) and social intent (a social exchange category); and both of those have greater relevance than the target character (*t*). The second encoding ($\texttt{glz}_{\texttt{intent}}$) puts the *intent* into the determinant and treats the initiator, social exchange, and target as equally important.

These encodings carry different design knowledge. In the former, it is assumed that every input is roughly comparable; in the latter, pursuing different social goals—improving friendship, beginning to date, becoming enemies—is treated as making fundamentally different maneuvers. If one encoding performs better than the other in creating play trace dissimiliarity comparisons that match the Prom Week designers' perception of dissimilarity, that tells us something about Prom Week: either social intent is one component of strategy among many, or else it is the primary indicator of player intention.

For a concrete example of each encoding, consider a turn in which the player wants Chloe to flirt with Doug. Here, the initiator is Chloe, the target character is Doug, the social intent is to increase Doug's romantic affection for Chloe, and the specific social exchange is flirting. In the $\texttt{glz}_{\texttt{ie>t}}$encoding, the input's determinant is simply `move` and the value contains `Chloe`, `Flirt`, and `Doug`, with `Doug` in a less-important position; in the $\texttt{glz}_{\texttt{intent}}$coding, the determinant is `romanceUp` and the value contains `Chloe`, `Flirt`, and `Doug` in equal prominence.

## Other Metrics

We want to evaluate Gamalyzer, but in order to do so we need a sound baseline. For this work, we compare Gamalyzer against three different baseline dissimilarity metrics. Each of these measures is grounded in previous analyses of Prom Week play data.

The first baseline is derived from a commonly used feature in play trace analysis: n-gram counting of social exchange action names (McCoy and Mateas 2012). Each play trace can be represented (lossily) as a vector of n-gram counts. We take the Manhattan distance (L1 norm) between those vectors and normalize it by the total number of n-grams appearing in both traces to obtain a number between 0 and 1. As in earlier work, the initiator and target characters were ignored in these n-gram counts, so this measure abstracts the play traces of interest.

The second baseline is inspired by unpublished work in clustering Prom Week play traces. Here, each play trace is represented as a vector of counts of interactions: an intent, an initiating character, and a target character. Many social exchanges can map onto the same intent, so this is also an abstraction of the original play trace. We take the normalized Manhattan distance between these vectors to yield a number between 0 and 1.

The first three metrics we have discussed—Gamalyzer, n-gram counting, and interaction counting—work on sequences of actions, although the latter two abstract that sequence into counts. Our final baseline is instead based on game *states*: the distance between two traces is defined as the distance between their terminal states. This state-based metric uses angular dissimilarity, where the feature vectors are comprised of the strengths of the three relationships (respect, platonic affection, and romance) between every pair of characters.

Gamalyzer and event counting are relatively game-independent, whereas the state distance metric is relatively game-specific (it happens to be easy to define for Prom Week). We might expect the more game-specific metric to perform better than interaction counting, which we would expect to beat n-gram counting because interaction counts include information about initiator and target characters.

From Gamalyzer's definition and underlying assumptions, if Gamalyzer does not beat the counting-based metrics and our first hypothesis is unsupported, then Prom Week's designers must perceive that making the same moves in a different order implies a similar strategy. If our experiments

do not support our second hypothesis—that is, if Gamalyzer does not beat the state-based metric—, Prom Week's designers must perceive that distinct sequences of moves indicate similar strategies (i.e. that the game has many symmetries).

In interviews conducted before the experiment, Prom Week's designers supposed that their game would not have a large number of symmetries. If our predictions about these metrics are correct, and if Gamalyzer does *not* outperform these baselines, that result could be viewed as evidence against the designers' belief in the game's lack of symmetries. This illustrates the importance of validating metrics experimentally, and of building well-founded theories from those observations. While this work on its own is insufficient to conclusively prove or disprove the hypotheses in the introduction, its contributions in connecting play trace metrics to characteristics of game dynamics should be of special interest to scholars of game design.

## Applying Dissimilarity Metrics

Distance measures do have an immediate utility for searching and filtering play traces, but they can also be applied to other purposes. In this paper, we consider one general-purpose application—finding outlier play traces—and another which is of special interest to Prom Week's designers (McCoy and Mateas 2012): describing the overall uniqueness of a set of play traces.

We also consider presenting a sorted list of outliers to a designer; this could highlight players who are misunderstanding a system, or who are playing to sabotage or circumvent it. We could also easily compare these ratings against human judgments to evaluate distance metrics. To determine the degree to which each play trace is an outlier, we need an operational definition of "outlier-ness" in terms of distances between play traces.

We derived a measurement using k-medoids, a classical partitioning technique. The medoid of a set of traces is the trace with minimum average distance to all the other traces in the set; to generalize to $k > 1$, we pick k elements (medoids) of the set so as to minimize the sum of the distances of each trace in the set to its nearest medoid. Informally, a medoid is like a centroid, except that it is not a mean, but a median (one of the elements in the original set). To calculate the degree to which a trace is an outlier with respect to a set of traces that contains it, we take its distance from the nearest medoid.

We can use this approach to judge the overall uniqueness of traces in a set: to a first approximation, we can imagine that the average outlier rating of the traces in the set is a proxy for the set's overall uniqueness. For sets where many traces are strong outliers, the uniqueness will be high, and for sets with few strong outliers, the uniqueness will be low.

## Experiment design

In our experiment design we take the dissimilarity of playtrace metrics as rated by one of Prom Week's designers as the ground truth against which we compare the metrics. The closer a metric gets to achieving the same dissimilarity ratings as the human designer, the better the metric is. The metrics used are the two different Gamalyzer encodings described in **About Gamalyzer** and the three non-Gamalyzer encodings outlined in **Other Metrics**. Several of our distance metrics involved parameters which had to be tuned (Gamalyzer's warp window $\omega$; k=1 or 2 medoids; and n for the n-gram metric). In each case, we used an automated search process to select parameter values that minimized root-mean-square error so that each metric (including the baselines) would be represented as well as possible. Our complete experimental setup including play traces, ratings, and analysis code are available at `https://github.com/JoeOsborn/metric-eval`.

Our experiments samples were drawn from the 3,186 complete play traces of Doug's story. To reach this point players must have built up a basic proficiency of manipulating the social space. Additionally, though Doug's scenario has multiple solutions to its social puzzles, its short length makes the designer's task of providing dissimilarity ratings tenable. Future work must examine whether some metrics are more or less appropriate in other levels of Prom Week. Individual traces consisting of a sequence of moves instigated by the player (an interactor, a social exchange, and a target character) were presented to raters as prose generated by a simple templating system which presented the characters involved and the intent of the social exchange.

Each of the ratings questions includes the language "with respect to player strategy." This is because play traces could be dissimilar in a variety of ways—with respect to player strategy, player experience, winning or losing, goals achieved, et cetera. We tried to frame the questions so that the raters would consider only the lens of player strategy (which seems the likeliest sort of similarity to derive using only player actions). By using three sets of rankings in three experiments, we demonstrate the accuracy of the metrics under test in answering a range of common game-design questions, giving a better sense of their overall utility.

The main concern in all three experiments is the low number of ratings relative to the population of traces. This is somewhat unavoidable, since most games have a small number of designers and a large number of play traces; it would be extremely difficult for so few people to annotate so many traces. Another issue compounded by the small number of raters is that a single rater might use different heuristics and internal criteria during different trials. If we had more raters, we could control for this; as it is, we have to assume that the designers have a good sense of play trace difference. Possible controls even when the number of raters is small include "warming up" each rater with several trials whose ratings will be discarded, or randomizing the order of the trials for each experience.

Though having additional raters might have mitigated the above dangers, we claim that using few raters is in and of itself not a threat to validity, but rather to generality: in other words, this work suffices to show validity for Prom Week, but not for other games. Comparing these metrics against ratings from more individuals would likely change the calculated errors, but those revised errors would be measuring a fundamentally different thing: a *group* of designers' perception of dissimilarity as opposed to that of a *single* designer.

Any noise or inconsistency in an individual designer's ratings are in fact part of the phenomenon we are trying to capture, since we hope to automatically approximate a game designer's perception of their own game's dynamics. Furthermore, considering that many games only have a single designer, the low number of raters is not in and of itself a threat to validity. We do feel that repeating these experiments with more designers of *different* games would show how the results of this paper generalize.

Finally, we must note that with only one rater working in a seven-point scale, we can't hope for any metric to have an error much less than $\frac{1}{7} = 0.143$. Greater precision than that on any individual trial would be difficult to justify.

### Experiment 1: Trace Dissimilarity

The fundamental question when evaluating a play trace dissimilarity metric is whether the metric is accurate. The natural experiment, then, is to compare the distances provided by some candidate metric against human-provided distances. For this experiment, we conducted 25 trials of the following scenario: we randomly selected a sample of six traces from the population, the first of which was designated as a reference; then, a distance was determined from the reference to each of the other five traces. Raters evaluated each trace's distance from the reference on a 7-point Likert scale ("On a scale of 1 to 7, with 1 meaning 'exactly the same' and 7 meaning 'incomparably different', how different is this trace from the reference trace with respect to player strategy?"). The ratings were normalized to a closed unit interval and compared against each of the distance metrics.

### Experiment 2: Outlier Rating

For our second experiment, we wanted to determine whether the Gamalyzer metric was the best choice of distance metric for outlier rating. We therefore needed designer ratings which described how much the designer perceived a given trace to be an outlier among a given set of traces. We conducted 25 trials in which 10 traces were randomly selected from the population. In this case, there was no reference trace; each trace was to be rated in terms of its "fit" with the rest of the sample. Raters determined this fit for each trace on a 7-point Likert scale ("On a scale of 1 to 7, with 1 meaning 'completely typical' and 7 meaning 'completely atypical', to what extent is this trace typical of this set with respect to player strategy?"). We normalized these ratings to a closed unit interval and compared them against the outlier rating measurement described above, using each of the underlying distance metrics.

### Experiment 3: Overall Uniqueness

Finally, we hoped to learn more about a core question underlying the evaluation of Prom Week as an interactive social AI system: Do players have unique experiences with the game? While our "player strategy" framing alters the tenor of this question somewhat, we can suppose that a player's choices are determined in large part by the player's experience of the game, and that their experience is also influenced by their choices. For this experiment, we conducted 25 trials in which 10 traces were randomly selected from the population. This experiment also used no reference trace. We asked raters to describe the whole set of 10 traces in terms of its incoherence—how unique or "spread out" the plays in this set were ("On a scale of 1 to 7, with 1 meaning 'completely uniform' and 7 meaning 'not at all similar', how similar are the traces in this set with respect to player strategy?"). These ratings were normalized to a closed unit interval and compared against the uniqueness rating measurement described above, using each of the underlying distance metrics.

| | Dissimilarity | Outliers | Uniqueness |
|---|---|---|---|
| $\texttt{glz}_{\texttt{intent}}$ | 0.208 | 0.279 | 0.189 |
| Interactions | 0.219 | 0.292 | 0.173 |
| 1-grams | 0.236 | 0.304 | 0.242 |
| $\texttt{glz}_{\texttt{ie>t}}$ | 0.287 | 0.326 | 0.290 |
| States | 0.592 | 0.557 | 0.576 |

Table 1: Root-mean-square error results for all metrics.

## Results and Discussion

Table 1 shows the root-mean-square error obtained between each of the five dissimilarity metrics and the game designer's ratings—ground truth—for each of the three experiments. The results seem to strongly support our second hypothesis: both encodings of Gamalyzer fared substantially better than the state based metric. The evidence for our first hypothesis is not quite as clear, since one Gamalyzer encoding was superior to the three baseline metrics while the other Gamalyzer encoding performed worse. Though clearly sensitive to the format of the input encoding, the fact that Gamalyzer *can* outperform other metrics is promising for its more widespread use as a game independent dissimilarity metric. But why did these metrics rank in this order? Answering this question could lead to improvements in Gamalyzer as well as new insights about *Prom Week*.

The trace dissimilarity experiment gives a foundational measure of suitability for a play trace distance metric. $\texttt{glz}_{\texttt{intent}}$(with $\omega$=20, the highest value possible for these experiments) narrowly edges out the interaction count metric and n-gram counting (n=1), but all three have error within two scale points of the human ratings. $\texttt{glz}_{\texttt{ie>t}}$($\omega$=7) fares slightly worse, while state distance performs badly.

The poor performance of state similarity might be because states describe outcomes and not strategies; the *mean error* of the state similarity metric is near -0.5, grossly underestimating dissimilarities (mean error is within $\pm 0.1$ for all other metrics). This behavior is consistent with two observations: many distinct sequences of actions might lead to similar states; and many similar sequences of actions might lead to different states (due to the hidden information and highly emergent dynamics of the game rules).

We believe that interaction counting beats n-gram counting by considering both the initiator and target of actions, and $\texttt{glz}_{\texttt{intent}}$improves over the interaction counting metric by accepting fuzzier matches and considering temporal ordering more strongly. But how do the event counting metrics get so close even though they consider much less informa-

tion? There must be temporal symmetries in the designers' perception of play trace differences. This came as a surprise to the designer who gave us the ratings, although it is unclear whether these symmetries are actually embedded in the game's dynamics or merely emerge from the designer's ratings. In the future, comparing perceived *trace* dissimilarities versus actual *state* dissimilarities could be used to help validate a game design.

N-gram counting performs only a little bit worse than interaction counting; why? It seems that within a given level (or at least within the scenario observed), the social exchange or social intent almost completely determines on its own the two agents involved. This is not to say that players do not have options; but once they have selected a social exchange, there is generally a small number of reasonable choices for the initiator and the target. This was a concern to one of the game's designers: was it possible that the opening narration of that scenario guided players too strongly? In other games or in other Prom Week levels (perhaps in a level with a variety of potential romantic interests), this determination might not hold. If the difference in error between n-gram counting and interaction counting did not increase in such levels, that would support the claim that Prom Week moves are largely determined by social exchange selection.

The substantial difference in performance between the two Gamalyzer encodings (and the good performance of the two counting metrics) shows that, although Gamalyzer is game-independent, the best choice of encoding varies from game to game. Gamalyzer encodings seem to perform better when the determinant (the *type* of the event) discriminates strongly in the same ways a designer would discriminate; otherwise, unrelated events will be perceived as more similar than they ought to be. In Prom Week, it appears that the strategic part of the move is the intent—that is, a *begin dating* move is so strongly different from a *become better friends* move that they cannot be compared directly. There is also a sizable difference in optimal warp window between the two encodings. The large warp window in $\texttt{glz}_{\texttt{intent}}$reflects the low temporal coherence required for designers to perceive similarity, while the smaller warp window in $\texttt{glz}_{\texttt{ie>t}}$is necessary to avoid underestimating dissimilarity; the interaction between game design, encoding, and warp window width is worth exploring in more detail.

Error rates for outlier rating and uniqueness appraisal are mostly in line with the trace dissimilarity experiment ($\mathrm{k} = 1$ medoid was chosen as it minimized error for all metrics). Strangely, although uniqueness is built on top of the definition of outlier rating, a uniqueness measure based on interaction counting beats one based on $\texttt{glz}_{\texttt{intent}}$—even though $\texttt{glz}_{\texttt{intent}}$outperforms interaction counting on the outlier rating task. Still, the difference in performance is much less than a single rating scale point.

Immediate future work includes reproducing these experiments on different Prom Week levels to test the degree to which intent determines initiator and target characters. Future experiments should use more ratings (perhaps increasing the size of each individual trial) to improve precision; it's also important to repeat these experiments in different games to test the claim that Gamalyzer is game-independent.

## Conclusions

The main contribution of this work is a methodology for evaluating play trace dissimilarity metrics, grounded in well-established techniques for evaluating other types of similarity measurements. In this experiment, two encodings of Gamalyzer were compared against three standard dissimilarity metrics; one state-centric measure and two action-centric measures. Both encodings performed better than the state-based metric, while only one encoding performed better than the counting-based metrics. We believe that this technique is widely applicable to other games and other metrics. Other operational definitions of play trace characteristics—questions that a good metric should help answer—could also be included in this instrument as appropriate to the game under consideration.

This work also builds evidence for Gamalyzer's claim of game-independent dissimilarity measurement, but future work must repeat these experiments on other games and against other baselines. Gamalyzer does seem to yield relatively accurate play trace dissimilarities, and it can be used effectively in derived measures. That said, it remains highly sensitive to the choice of input encoding, and our findings in this paper do imply that certain encodings lend themselves better to certain games. We also suspect that other design perspectives besides "player strategy" would be better served with specialized encodings. The space of reasonable encodings is relatively small for a given game, so it is feasible to find the best encoding through experimentation; but it would still be helpful to know more about the interaction between the game design and choice of encoding.

In order to be appropriate for broader use, Gamalyzer's documentation must provide clearer guidelines on what makes an effective encoding; moreover, tools should be developed that can guess at an encoding's quality based on properties like the number of parameters in each event, the number of distinct determinant types, and so on. As Gamalyzer matures and is used (and validated) in more games, the characteristics of good and bad Gamalyzer encodings will become more apparent, allowing us to provide more guidance about playtrace encodings to future Gamalyzer users.

If we can strongly validate a game-independent play trace dissimilarity metric against the intuition of professional game designers, new categories of general game design support tools will be possible. This will involve answering questions such as "Which metrics (or families of metrics) are most effective for which games?" The idea that some metrics are better or worse for certain games is also fascinating: if this is due to hidden properties of a given game's design or dynamics, we might be able to use the appropriateness of a metric as a proxy for those hard-to-measure properties and evolve our understanding of the science of game design.

## Acknowledgements

Special thanks are due to Sri Kurniawan for her advice and feedback on earlier drafts of our experiment design, as well as to Prom Week's players and designers.

## References


Andersen, E.; Liu, Y.-E.; Apter, E.; Boucher-Genesse, F.; and Popović, Z. 2010. Gameplay analysis through state projection. In *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, 1–8. New York, NY, USA: ACM.

el Nasr, M. S.; Drachen, A.; and Canossa, A. 2013. *Game Analytics*. Maximizing the Value of Player Data. Springer.

Kim, J. H.; Gunn, D. V.; Schuh, E.; Phillips, B.; Pagulayan, R. J.; and Wixon, D. 2008. Tracking real-time user experience (true): a comprehensive instrumentation solution for complex systems. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 443–452. ACM.

McCoy, J. A., and Mateas, M. 2012. *All the world's a stage: a playable model of social interaction inspired by dramaturgical analysis*. Ph.D. Dissertation, University of California at Santa Cruz.

McCoy, J.; Treanor, M.; Samuel, B.; Reed, A. A.; Mateas, M.; and Wardrip-Fruin, N. 2013. Prom week: Designing past the game/story dilemma. In *Proceedings of the Eighth International Conference on the Foundations of Digital Games*.

McCoy, J.; Treanor, M.; Samuel, B.; Reed, A.; Mateas, M.; and Wardrip-Fruin, N. 2014. Social story worlds with comme il faut. *IEEE Transactions on Computational Intelligence and AI in Games* PP(99):1–1.

Osborn, J. C., and Mateas, M. 2014. A game-independent play trace dissimilarity metric. In *Proceedings of the Ninth International Conference on the Foundations of Digital Games*.

Rogowitz, B. E.; Frese, T.; Smith, J. R.; Bouman, C. A.; and Kalin, E. B. 1998. Perceptual image similarity experiments. In *Photonics West'98 Electronic Imaging*, 576–590. International Society for Optics and Photonics.

Russell, R., and Sinha, P. 2011. A perceptually based comparison of image similarity metrics. *Perception* 40(11).

Samuel, B.; McCoy, J. A.; Treanor, M.; Reed, A.; Wardrip-Fruin, N.; and Mateas, M. 2014. Story sampling: A new approach to evaluating and authoring interactive narrative. In *Proceedings of the Ninth International Conference on the Foundations of Digital Games*.