Crowdsourcing Open Interactive Narrative

Matthew Guzdial, Brent Harrison, Boyang Li, Mark O. Riedl
School of Interactive Computing
Georgia Institute of Technology
{mguzdial3; brent.harrison; boyangli; riedl}@gatech.edu

ABSTRACT
Interactive narrative is a form of digital interactive experience in which users influence a dramatic storyline through their actions. Artificial intelligence approaches to interactive narrative use a domain model to determine how the narrative should unfold based on user actions. However, domain models for interactive narrative require artificial intelligence and knowledge representation expertise. We present open interactive narrative, the problem of generating an interactive narrative experience about any possible topic. We present an open interactive narrative system—Scherazade IF—that learns a domain model from crowdsourced example stories so that the player can perform different actions and still receive a coherent story experience. We report on an evaluation of our system showing near-human level authoring.

Categories and Subject Descriptors
I.2.1 [Artificial Intelligence]: Applications and Expert Systems—Games

General Terms
Algorithms, Human Factors

Keywords
Procedurally generated games, interactive fiction, crowdsourcing

1. INTRODUCTION
Creating video games requires a great deal of authorial input and specialized knowledge. The reliance on human-authored content has implications for serious games in particular, which delays the development of new educational and training experiences. Specific genres, such as *interactive narrative* rely heavily on content production due to the branching nature of the experiences they offer.

Interactive narrative is a form of digital interactive experience in which users create or influence a dramatic storyline through their actions, either by assuming the role of a character in a fictional virtual world, issuing commands to computer-controlled characters, or directly manipulating the fictional world state [14]. The goal of interactive narrative is to immerse the user in a virtual world such that he or she believes that they are an integral part of an unfolding story and that his or her actions have meaningful consequences. The simplest interactive narratives are branching stories, such as Choose-Your-Own-Adventure books and hypertexts in which each plot point is followed by a number of options that lead to different, alternative narratives unfolding. As the number of paths grows exponentially with each new plot unit, authoring new plot units that are compatible with each possible path becomes increasingly difficult.

More complex interactive narrative systems use artificial intelligence (AI) to determine available options to the user. One application of AI techniques to interactive narrative attempts to overcome the combinatorics of authoring branching stories. Intelligent interactive narrative systems use knowledge about the fictional story world to automatically determine which options should be presented to the user and how the user's narrative experience should subsequently unfold. Common approaches to intelligent interactive narrative include search-based drama management [9, 21], planning [7, 16], case-based reasoning [17], and machine learning [11, 22]. Intelligent interactive narrative vastly reduces the authorial burden of creating interactive narrative experiences by trading the authoring of explicit authoring of narrative branches for the authoring of *domain models* that compactly express possible future narrative trajectories plus criteria to evaluate possible future narratives. However, the creation of domain models can be difficult and require expertise in artificial intelligence and knowledge representation.

We introduce *open interactive narrative*, the problem of generating an interactive narrative experience about any possible topic. Unlike prior approaches to intelligent interactive narrative, an open interactive narrative system learns the domain model from which it constructs an interactive experience. Prior approaches to intelligent interactive narrative can only generate stories involving the content encoded into a given domain model. Automatically learning the domain model has the added benefit of further reducing the authorial burden; authors using open interactive narrative systems do not need expertise in programming or encoding knowledge in an AI representation. In theory one only needs to tell an open narrative intelligence system what one wants the narrative experience to be about to generate interactive narratives of human authored quality.

We present an intelligent system, Scheherazade-IF, an open interactive narrative system that automatically generates interactive fictions about common topics. Scheherazade-IF uses crowd-sourcing to learn a domain model for a given topic in a just-in-time fashion, meaning it engages in domain knowledge acquisition during development of the interactive narrative experience about a specific, requested topic. Crowdsourcing is the outsourcing of complicated tasks—typically tasks that AI cannot perform well by itself—to a large number of anonymous workers via Web services [13]. Scheherazade-IF delegates the authoring of domain knowledge to a large number of anonymous workers. We do not assume crowd workers possess expertise in computer science, modeling, or interactive narrative. Instead, crowd workers are asked to provide linear archetypical examples of what could happen during the interactive narrative in natural language for the given topic; this is a natural mode of communication for humans. This yields a highly specialized corpus of example

narratives from which a general model of the topic, known as a *plot graph*, can be learned.

The plot graph representation has been previously used in interactive narratives [9, 21], but the specific plot graphs were hand-authored. The work presented in this paper shows that plot graphs can also be learned automatically in open interactive narrative. Previously we made use of this same plot graph learning approach in non-interactive story generation [5]. In this paper we demonstrate its extendibility into interactive domains. This addition of interactivity required addressing issues of player autonomy and non-player character (NPC) behavior.

To the best of our knowledge, this is the first attempt to automatically generate an executable interactive experience without reliance on a handcrafted domain model. The contributions of our work are as follows: (1) the techniques used to make plot graphs derived from crowdsourced information playable, and (2) the results of a successful evaluation of Scheherazade-IF against human authored interactive narrative.

2. RELATED WORK

Common approaches to intelligent interactive narrative include search-based drama management [9, 21], planning [7, 13, 16], case-based reasoning [17], and machine learning [11, 22]. Many of these approaches use a *Drama Manager* (DM), an autonomous, omniscient, non-embodied agent that attempts to maximize a set of author-provided heuristic functions to improve user experiences.

Search-based Drama Management [9, 21] uses adversarial search to select DM actions—causers, deniers, and hints— that increase the likelihood that the player will follow a trajectory that scores well. Declarative Optimization-based Drama Management builds on this but uses reinforcement learning to account for uncertainty of player actions when selecting DM-actions [11].

Search-based Drama Management and Declarative Optimization-based Drama Management encode domain knowledge as a *plot graph*. A plot graph is a temporally ordered model that determines the logical flow of events in a fictional world [22]. A plot graph is a directed acyclic graph in which nodes are plot points—major events and actions of players and non-player characters—and edges indicate temporal precedence relations between plot points. For example, finding the vault and its key must precede opening the vault. Nelson and Mateas extended the original plot graph representation with OR-relations between plot points, indicating that a plot point can be reached by a variety of distinct means [9].

Giannatos et al. describe a technique by which a genetic algorithm modifies a plot graph by suggesting new plot points and new precedence constraints that prune undesirable narrative sequences [2]. However, the technique cannot produce semantic interpretation of new plot events; it can only determine that there should be another plot point in a particular place in the graph.

The above techniques for intelligent interactive narrative require a domain model, such as a plot graph, to be authored before it can guide player experiences. The work presented in this paper is a technique for automatically producing an interactive narrative experience from scratch. The automatic generation of a playable interactive experience falls into the category of *procedural game generation*. Unlike procedural content generation such as level generation, procedural game generation attempts to produce all aspects of a playable experience. Togelius and Schmidhuber generate "pacman-esque" grid based games by using a genetic

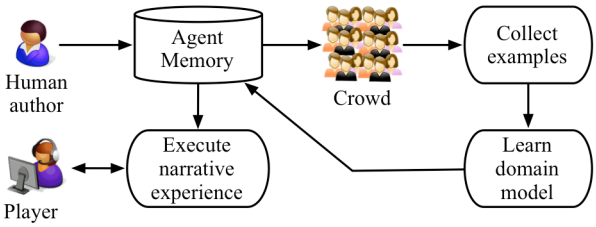

Figure 1. The Scheherazade-IF Architecture

algorithm to permute basic gameplay axioms about win states, rules, initial placement of game elements, and various other requirements for a game system [20]. The system is bootstrapped by pre-existing gameplay axioms (e.g., elements have location and can move and collide) and a fitness function based on learning progress.

Cook et al. introduce the *Mechanic Miner* system, which generates novel movement mechanics for platformer games via a reflection-driven generation technique and hand written rules [1]. Zook and Riedl generate novel mechanics across various genres by structuring mechanic generation as a constraint satisfaction problem [24]. Both systems assume the inclusion of a human designer to craft design requirements at some stage in their generation. This limits these approaches in terms of the knowledge barrier required to use them.

Nelson and Mateas [10] and Treanor et al. [21] describe techniques for generating games in the style of the *WarioWare* series. The former make use of large-scale corpora—specifically WordNet [12]— to identify semantically related concepts. The *Say Anything* system makes use of thousands of weblogs in a collaborative storytelling game between the system and a human player [19]. Instead of working off pre-compiled general-purpose corpora, our work uses crowdsourcing to collect a highly specialized corpus of narrative examples from which to mine knowledge about plot points and their temporal precedence relations.

We are not alone in making use of crowdsourcing to inform game generation. *The Restaurant Game* is a system that crowdsources interactions between individuals in a typical restaurant [12]. Unlike our own approach, The Restaurant Game has a fixed, underlying domain model—the types of actions are known, but not the orderings. Sina et al. inform a training system with crowdsourced, semi-structured stories to serve as alibis for virtual suspects [18]. Alibis are presented to users but are not themselves playable experiences. Additionally, both make use of crowdsourcing models to inform a specific experience, whereas our approach allows us to construct a wide range of experiences (e.g. bank robbery, movie date, etc.).

3. PLOT GRAPH LEARNING

We present a brief overview of the Scheherazade-IF architecture in this section, but see [4, 5] for more information about the underlying story generation system that has previously been utilized for non-interactive story generation. The underlying representation of story knowledge in Scheherazade-IF is that of a *plot graph*. The plot graph serves as a guide that Scheherazade-IF follows for how to construct an interactive narrative experience for a human player. Scheherazade-IF is a just-in-time learning system, meaning that if the system is unfamiliar with the domain of the interactive narrative experience to be generated, it must first

learn a new domain model. Figure 1 shows the system's architecture.

Humans interact with the system in one of three ways, as either players, members of the crowd, or "authors". Authors request Scheherazade-IF to construct a novel interactive narrative about a specific subject or situation. If Scheherazade-IF does not have a model of the situation in its memory, it seeks to learn about the situation from the crowd. The crowd informs Scheherazade-IF with example stories concerning the requested situation. Scheherazade-IF compiles these examples into a domain model in the form of a *plot graph*. The compilation of examples into a model is a key step that allows interactive execution; the stories are woven together so that at any step in time the player will have alternatives to choose from and sensible consequences. Scheherazade-IF has an understanding of the dynamics of the situation and is not just a collection of stories.

Plot graphs represent Scheherazade's understanding of the dynamics of a narrative situation (e.g. a movie date). In our work, a plot graph G is a tuple (E, P, M, E_o, E_c). E is the set of events. $P \subseteq E \times E$ is a set of precedence relations between events. Mutual exclusions relations between the events belong to the set $M \subseteq E \times E$. Finally, $E_o \subset E$ is a set of optional events and $E_c \subset E$ is a set of events that are conditioned on the optional events. In the graphical representation, each event is presented as a node in a graph. For an example see Figure 3, a plot graph Scheherazade-IF generates for the situation "movie date". The arrows between nodes are precedence relationships between events. These relations serve to encode typical orderings of events found within situations and their temporal nature. The mutual exclusions—the dashed lines between certain nodes—indicate that certain nodes cannot co-occur in any story or interactive experience that results from executing the plot graph. Mutual exclusions enforce coherence within an interactive narrative, and allow for branching options. For example in Figure 3 if one meets Sally at the theater, one should not also drive her home (as is assumed in the "go home" event, as she would have driven her own car). The mutual exclusion enforces the effects of the action, though Scheherazade-IF has no direct understandings of actions or effects.

Figure 2 shows two possible beginning of two non-interactive stories the "movie date" plot graph can generate. As with all the plot graphs in Scheherazade-IF's memory, the plot graph in

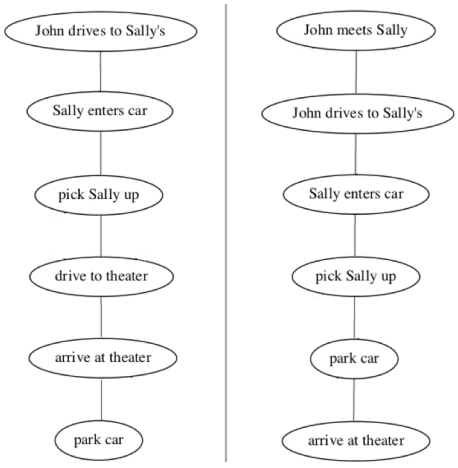

Figure 2. Two separate linear traces through the "movie date" plot graph.

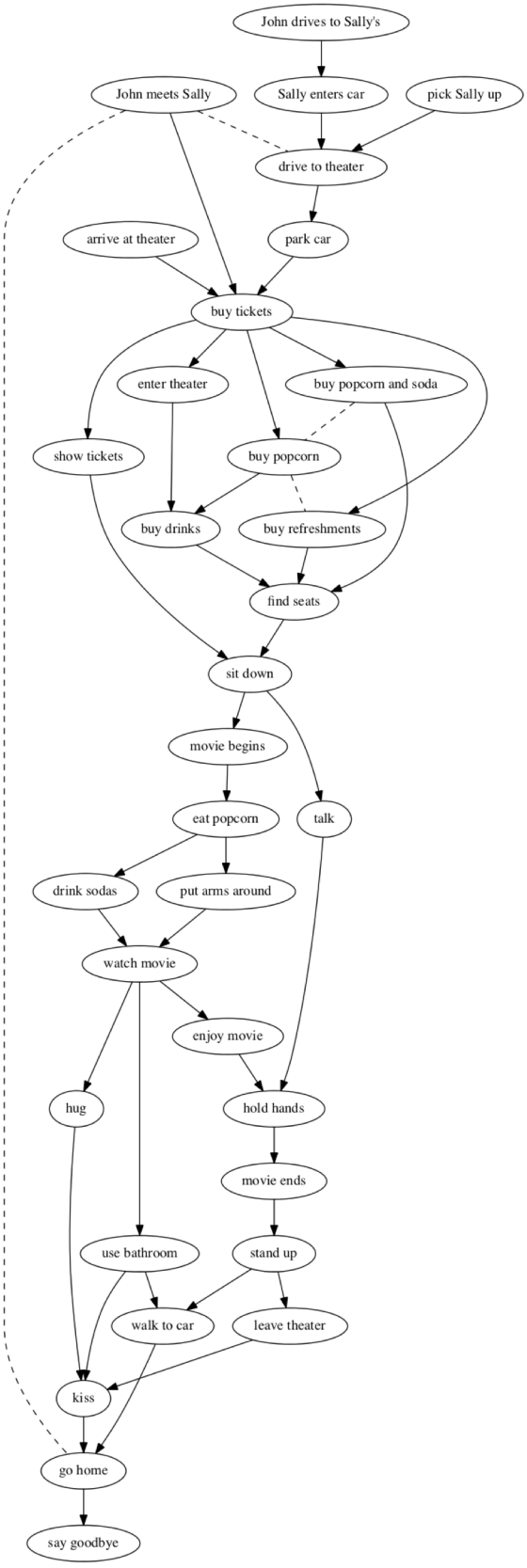

Figure 3. A representation of the movie date plot graph.

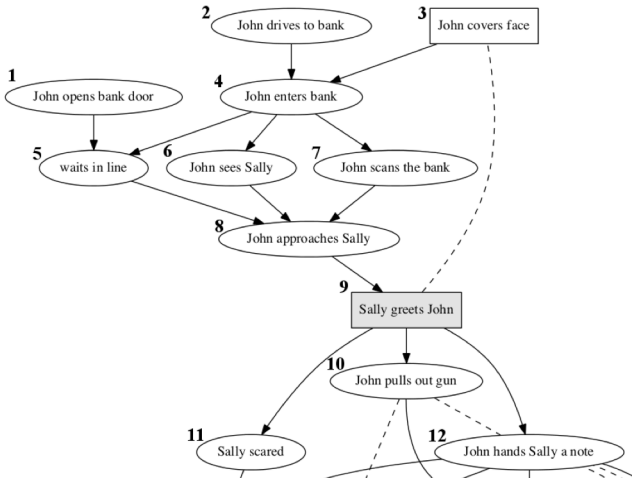

Figure 4. A subsection of the bank robbery plot graph.

Figure 3 represents the space of stories for a specific domain. ("movie date" in this case). Multiple stories can be generated via walking through the plot graph according to its structure. To begin with, the only nodes available are the three without precedent relationships. Including the "John meets Sally" node in the right side means that the option to "drive to the theater" should not be presented and thus does not appear in the second trace.

Scheherazade-IF acquires plot graphs from raw crowdsourced natural language data from Amazon Mechanical Turk (AMT), a crowdsourcing platform. The task requires AMT's anonymous workers to input simple linear stories of typical ways in which the given situation can unfold. To simplify the problem of natural language processing, the crowd workers are asked to simplify their language in three ways: make use of specific characters for certain roles in the story, segment the narrative into events such that each sentence contains a single activity, and to write as simply as possible with one verb per sentence.

Once example stories are collected, Scheherazade-IF must do two things to compile a plot graph: (1) it must determine what the primitive events are for the given situation, and (2) it must determine the structure of the plot graph by identifying typical event orderings and mutual exclusions. Note that Scheherazade-IF has no *a priori* knowledge about what actions people can take in different situations—this must be learned. Primitive events are clusters of sentences from different examples that semantically refer to the same activity. The Stanford parser is used to extract syntactic dependencies from individual sentences [3]. WordNet is used to calculate the semantic distance between sentences based on the individual words used in that sentence and how they align with parts of speech in other sentences [8]. The intuition here is that sentences that appear consistently represent archetypical events within a specific set of stories, and thus should be represented in the plot graph.

From the events, Scheherazade-IF generates a plot graph's structure by identifying precedence relations and mutual exclusions. The system recognizes precedence relations by determining whether the probabilistic confidence that a given event in a story precedes another based on the ordering of associated sentences in the crowdsourced examples. The graph construction process is formalized as an integer quadratically constrained problem, which serves to avoid cycles while keeping the most probable precedence relations possible. Lastly the system identifies mutual exclusions between events. Similar to the

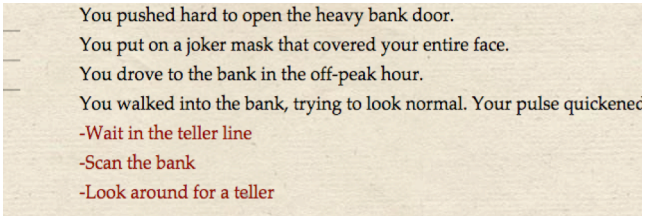

Figure 5. A screen shot of the Scheherazade-IF being played.

precedence relations, it determines this relationship based on the probability that an event does not exist within any crowdsourced stories when another is present. Optional and conditional events are identified as events that are mutually exclusive with their own descendants in the graph—the ancestor is marked optional and the descendant is conditional on the first. In the graph, these events are represented as box nodes, as seen in Figure 4.

The quality of the generated plot graph in terms of coherence and lack of commonsense errors depends upon the coverage of the data. By coverage, we mean the number of stories that semantically cover the same events. For example, the "movie date" stories from the AMT crowd workers begin at different stages of the situation, with some stories starting at home and others starting after arrival at the theatre. This data sparsity at the beginning leads to a high degree of commonsense errors, as can be seen from Figure 3. In this plot graph, it is possible to "meet Sally" at the theatre, only to have to still "drive to Sally's" to continue. In this paper, we seek to understand how these commonsense errors impact the quality of the experience of the interactive player.

4. INTERACTIVE STORY GENERATION

Plot graphs are models of situations, as opposed to specific narratives. To make an interactive narrative from a plot graph, the plot graph must become playable. Previously, we have made use of random walk search as a method for traversal for non-interactive story generation, as plot graphs enforce coherence. However, with a player, that method of traversal no longer applies. The initial problem then is how does Scheherazade-IF decide what actions/events to present to a player at any point?

To present choices to the player the system first determines which plot events are *executable*. An event is executable when all of its direct, non-optional predecessors have been executed, except those parents excluded by mutual exclusion relations. For example, at the beginning of the "bank robbery" plot graph in Figure 4, if the player were playing as John, the executable events would be 1, 2, and 3 as they are the only events without predecessors and would then be presented to the player. The player then could choose the plot event 3 "John covers face". Choosing this option sends a message to the system to send back a text description of the event, which is then displayed to the player.

Once a plot event from the list of executable events is executed, it becomes part of the history, and Scheherazade-IF removes any event mutually exclusive with the executed event and recursively removes any event temporally dependent on an event already removed. Therefore choosing plot event 3 leads to the removal of event 9 "Sally greets John". Because optional events can be skipped, Scheherazade-IF also removes any optional events for which their temporal descendants have been executed. For example, if instead of choosing event 3, if the player had chosen events 1, 2, and 4 then event 3 would have been removed. To avoid losing structural information, direct parents of removed

events are linked to direct successors of removed events with temporal precedence links. In our example, due to choosing event 3, event 9 is removed. To retain structural information, events 10, 11, and 12 now have 8 as a precedent link.

Continuing with the example, if the player chose event 3 to begin with, then the system would present events 1 and 2 to them next. From the plot graph it may appear that the next available event from 3 would be event 4 due to the precedent link arrows. However, since the current executable events are dependent on their precedent relationships having been executed, event 4 only becomes available after events 2 and 3 are in the history.

In Scheherazade-IF the player can choose to play any of the characters present in the interactive narrative. These include "John" and "Sally", the characters we explicitly asked AMT crowd workers to include in their stories. However, the subject of any sentence that occurs frequently enough becomes an additional character. For example, "the police" appear later in the robbery plot graph. We found that in general though, a central protagonist emerges in any plot graph, both in terms of the number of actions and the agency that character demonstrates. In the case of both the bank robbery and movie date plot graphs this protagonist character was "John". Even though Sally has a major minority of plot graphs in the robbery plot graph, John has all of the actions that advance the story, while Sally's actions are largely reactive.

Players are presented with executable actions that are performed by the characters selected by the player. Once the player makes a choice, that action is marked as executed, and Scheherazade-IF re-computes the executable events. Depending on the plot graph it is possible that executable events may need to be performed by characters other than the one controlled by the player. The system handles the characters not chosen by the player as non-player characters (NPC). If the there are no NPC events in the set of executable events, Scheherazade-IF waits for the player to make a choice, as above. If there are no player actions in the set of executable events, Scheherazade-IF randomly chooses an NPC action, displays it to the player, and marks it as executed. Thus it is possible for the player to see events appear spontaneously although the player has taken no action. This distinguishes Scheherazade-IF from Choose-Your-Own-Adventure branching stories—the plot graph is a story-based simulation in which NPCs can act when necessary. Often the set of executable events has a mix of player and NPC actions. In this case, the system waits for a predetermined amount of time (currently set for 5 seconds based on informal play testing) and then randomly chooses an NPC action. This creates a race condition at times between player and NPCs, which creates a more dynamic and uncertain experience. For example, in the bank robbery game, there are two branches near the end resulting in either the player getting away or the player being caught by the police. After a certain point, if the player is too slow to take actions, the police will arrive and the game will end with the player being caught.

To present events to the player, Scheherazade-IF only has access to the simple sentences that are clustered into events during plot graph learning. The simplicity of these sentences is valuable in helping Scheherazade-IF learn a domain, but does not make for compelling reading in an interactive fiction. Additionally, the sentences are not written in the second person. See [6] for work on natural language generation that involves crowdsourcing more interesting descriptive sentences to generate text with different narrative styles. Rewriting sentences in second person is future work. For purposes of evaluating how well Scheherazade-IF translates the plot graphs to interactive narratives, the current version of Scheherazade-IF uses manually written descriptive sentences for each event.

5. EVALUATION

Our system is designed to produce interactive experiences comparable to those crafted by a human author. Toward that end, we evaluated Scheherazade-IF against two baseline interactive narrative generators, representing the theoretical most and least intelligent versions of the system. For the upper bound we made use of a human expert to construct the best plot graph possible. For the lower bound, we made use of a plot graph with random temporal relations and mutual exclusion relations. In order to draw conclusions, all three versions of the system made use of the same events but with different precedence and mutual exclusion relations. The relations are an important part in determining the quality of possible experiences that can unfold. The purpose of the evaluation is to pinpoint the quality of experiences that can be produced by a fully automated interactive narrative generation system that works from noisy, stochastic crowdsourced data. Perceived errors in Scheherazade-IF are typically due to the system missing precedence relations or mutual exclusion relations due to its lack of commonsense knowledge, allowing for events to appear out of place or to occur in a nonsensical order. In all three conditions, we wished to understand how errors affected the quality of players' experiences and how aware players were of the errors within the plot graphs. An example of an error stemming from the system's lack of commonsense knowledge would be the ability to open the bank door after already having entered the bank. Results from the study tell us how much better Scheherazade-IF is from a random baseline and how close it is to achieving human-level quality. Scheherazade-IF's abilities will be plotted as a point between that of the random baseline and the human baseline, which we refer to as the *R-H value*.

5.1 Baselines

The Human baseline is an upper bound of what we can expect an interactive narrative generation system to achieve. The Human baseline is a version of Scheherazade-IF that uses a plot graph manually constructed by an expert. Due to the constraint of making use of the same plot events, we created the human-authored graph by "correcting" graphs initially generated by Scheherazade-IF. We accomplished this by adding relations that appeared to be missing and by removing relations that appeared incorrect. The expert, the first author of this paper, iteratively tested his modifications to the graph until he could not improve the quality of the story experiences that could be generated. To avoid bias, we had other experts in interactive narrative who were not affiliated with the project play through the interactive narratives generated from these "corrected" plot graphs and made any suggested changes.

The Random baseline is a lower bound on what we might expect from an automated interactive narrative generator. We adopted the following procedure for generating random plot graphs. We used the set of events initially generated by Scheherazade-IF, but discarded all temporal precedence relations and mutual exclusion relations. We then randomly added precedence relations and mutual exclusion relations that did not create cycles. We stopped when the plot graph was capable of producing story experiences of the same length as the plot graph learned by Scheherazade-IF. This check is essential to ensure that trivial one- or two-step stories would not occur.

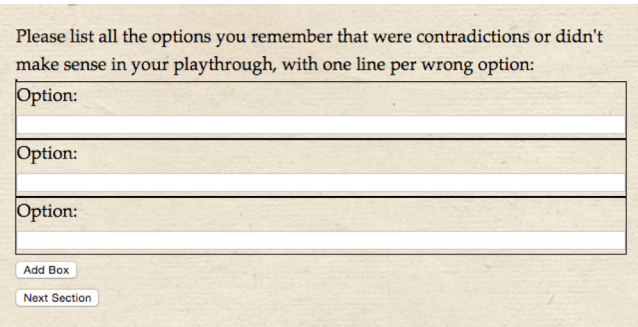

Figure 6. A screen shot of the recalled errors section.

5.2 Methodology

We undertook a between-subjects factorial study in which each individual played through interactive narratives generated from a single interactive narrative generator—the full Scheherazade-IF system, the Human baseline system, or the Random baseline system. Each participant played through two scenarios—a bank robbery and a date at a movie theatre. The order in which the scenarios were presented to participants was randomized to avoid bias from ordering. We made use of two different situations as a means of demonstrating generalizability of the generator. After playing through each interactive narrative the subject would answer a number of quantitative and qualitative questions. For the quantitative sections, we made use of commonsense errors as a means to evaluate similarity to an ideal human author. We have made use of this metric previously [4], as it allows statements concerning how well a generator mimics the knowledge humans gain via everyday life. For the quantitative sections we made use of story understanding, enjoyment, story type determination, and agency as these are all important characteristics of the interactive narrative media [22].

After each playthrough, we required individuals to fill out a series of self-report measures on a Likert scale from 1 to 6. The questions were as follows:

1. I understood the story of the interactive narrative. (Understanding)
2. The story largely made sense. (Understanding)
3. I enjoyed the story of the interactive narrative. (Enjoyment)
4. I couldn't determine what type of story this was. (Story type determination)
5. I didn't understand what was happening in the story. (Understanding)
6. I felt I could take an active role in the story. (Agency)

We made use of three questions to report understanding as coherence was a primary concern of ours. Question five was used as a means of gauging whether participants on Amazon Mechanical Turk were paying attention.

To measure the effect of plot graph learning errors on players, we measured the number of errors that players observed. Errors were measured in two different ways. First, we asked participants to write down all the places in the interactive narrative experience that they remembered seeing errors. The intuition behind this measure is that only some errors will be significant enough to be recalled without prompting (a recall task). Participants were asked to write down lines in the story that were "contradictions or did not make sense in your playthrough." See Figure 6 for the user interface for eliciting recalled errors. If an individual attempted to move on with entirely empty text entry boxes, a popup asked them to put "Not Any" into the first box. Additionally, while we included the "Add Box" button to allow for another input box to be added manually, typing anything within the last open box would always add another text input box.

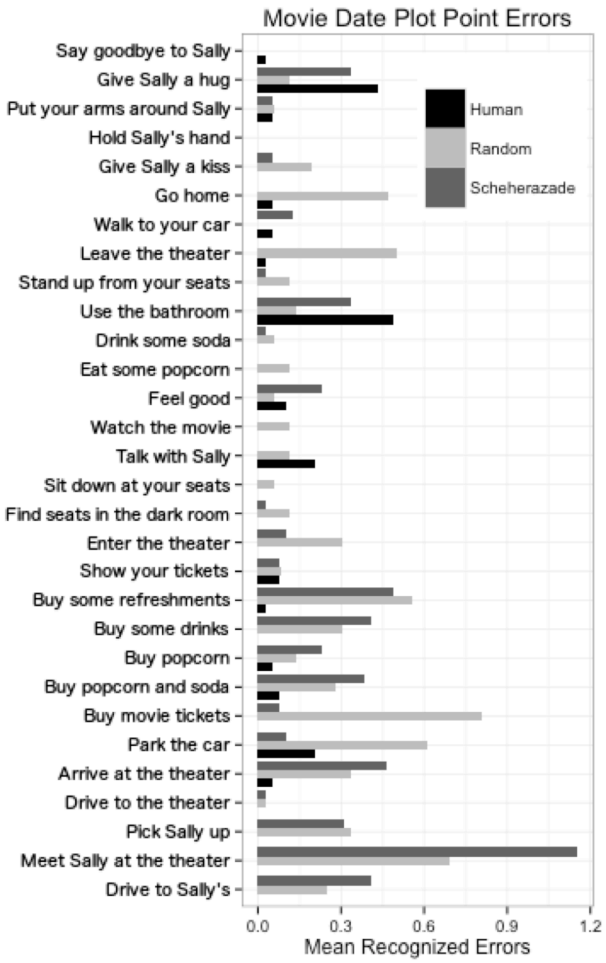

Figure 7. A graph of the mean recognized error values for each plot point for the movie date interactive narrative

Our second measure of perception of errors during interactive play is after-the-fact recognition of errors (a recognition task). In this task, participants were shown a trace of the events and options that they experienced during their play session. Participants were asked to mark all of the errors, using the same criteria as before. A checkbox would appear next to each line of the trace.

We ran the study on Amazon's Mechanical Turk. The study was projected to last 15 minutes, and each participant was paid $2.50.

5.3 Results

In total we collected 120 complete responses, 40 subjects for each of the three plot graph generators. We recruited from Amazon Mechanical Turk, which suggests some degree of Internet literacy and ability to read/write English in our population. Six participants reported some variation of "all were errors" in the recalled errors section—four from the random condition, and one from each of the other conditions. We threw out these responses; we believe these were attempts to game the task and get paid without work. Additionally there were twelve cases in which individuals recalled errors but then reported recognizing none. Participants either misremembered or adjusted their responses

Table 1. Summarizing the results of recognition errors by generator type

	Robbery	Movie Date	Modified Movie Date
Human Median	3	3	3
Scheherazade Median	3	10	5
Random Median	12.5	18.5	15
R-H Value	100%	54.8%	83.3%

after learning more about what to expect from the system; two-thirds of the times it occurred, it happened in a participant's second play through. We included these data points in our results.

The mean number of errors recalled in all three conditions was 2.2. We found a ceiling effect at three errors recalled, likely caused by the user interface (Figure 6), which started with three text-entry boxes, but expanded on need. It is possible that the three boxes primed participants to only think of three errors. Given that the random condition generates many more errors, we consider this data unreliable and do not consider it further.

Error recognition results are shown in Table 1. Note that we did not see the same ceiling effect with error recognition as with recalled errors as subjects merely had to identify moments that contained commonsense errors from a print out of their playthrough. We report the median number of recognized errors for each condition and for the two scenarios. We made use of median values as they do a better job of accounting for outliers. The R-H value is the performance of the full AI of Scheherazade-IF interpolated as a point between the performance of the Random

$$RHValue = 1 - \frac{(ScheherazadeMedian - HumanMedian)}{(RandomMedian - HumanMedian)}$$

condition and the Human condition, such that a 100% R-H value indicates a result identical to human results. Formally,
The robbery interactive narrative is identical to human in terms of recognized errors. For the raw movie date scenario, recognized error results gave an R-H value of about 54.8%. We believe a significant number of errors are due to sparsity of data at the beginning of the scenario. As noted earlier, some crowdsourced examples started at Sally's home and others started at the theater, resulting in less overall confidence about the temporal relations among the first few event nodes. Removing the effected nodes—specifically, the first four events of the movie date interactive narrative from those counted in the recognition data—we compute an R-H value increase to 83.3% from 54.8%. We believe this value is indicative of Scheherazade-IF performance when provided with sufficient data for all aspects of a scenario. To test our assertion, we removed four events at random ten times and found that this lead to an R-H value increase to 56.5% on average from 54.8%. Thus, the recognized errors are disproportionately due to the first four events in the movie date plot graph, as shown in Figure 7.

Results involving participants' subjective interpretations of the scenarios they experienced followed a similar pattern. Questions probed participant's perceptions of understanding, enjoyment, agency, and story type recognition. For the robbery scenario, there is no significant difference between Scheherazade-IF and the Human condition using the Wilcoxon-Mann Whitney Test as summarized in Table 2. There was a significant difference between Human and Random and between Scheherazade-IF and Random.

Table 2. Table of Wilcoxon-Mann Whitney Test p-values between each generator for the Bank Robbery.

	Q1	Q2	Q3	Q4	Q6	Q7
R-S	**$2e^{-11}$**	**$1e^{-13}$**	**$3e^{-7}$**	**$2e^{-4}$**	**$2e^{-6}$**	**$3e^{-7}$**
S-H	0.9	0.8	0.6	0.2	0.7	0.7
H-R	**$9e^{-13}$**	**$3e^{-14}$**	**$7e^{-6}$**	**$2e^{-3}$**	**$7e^{-8}$**	**$2e^{-6}$**

Table 3. Table of Wilcoxon-Mann Whitney Test p-values between each generator for the Movie Date.

	Q1	Q2	Q3	Q4	Q6	Q7
R-S	**$5e^{-6}$**	**$3e^{-7}$**	**$2e^{-5}$**	0.1	**$4e^{-4}$**	**$4e^{-6}$**
S-H	**$6e^{-6}$**	**$6e^{-5}$**	0.2	0.07	**0.03**	**0.04**
H-R	**$1e^{-11}$**	**$6e^{-15}$**	**$8e^{-8}$**	**$3e^{-4}$**	**$9e^{-8}$**	**$7e^{-10}$**

As summarized in Table 3, for the movie date scenario, Scheherazade-IF was more similar in entertainment value to the Human condition but more similar to Random in terms of recognized story type. For self-reports of understanding and agency, Scheherazade-IF was significantly different from both Human and Random conditions. This fit with the pattern that for the movie date scenario, Scheherazade-IF (without the removal of the sparse data points) falls halfway between Random and Human.

All participants experienced both scenarios back-to-back. Between the first and second interactive narrative sessions, subjects reported 20% fewer errors, measured both in terms of median values and total sums. We suspect this was due to questionnaire fatigue, as the second recognizing-errors section took place at the very end of the study.

We asked individuals to rank their familiarity with fiction reading, video games, Interactive Fiction, and choose your own adventure books. Though we found no connection between these self-reported values and how well individuals did by any metric, we did find that participants in the Human condition consistently ranked themselves differently in expertise on fiction reading and video games according to the Wilcoxon-Mann Whitney Test ($p > 0.05$). A median values analysis indicate that individuals ranked themselves higher in these two areas when given the human generator interactive narrative stories. We hypothesize that this phenomenon was due crowd workers having to work harder to find the commonsense errors in the human interactive narratives. These issues stand as potential issues for future evaluations of crowdsourced, creative generative systems.

5.4 Discussion

Both the quantitative error-recognition results and the qualitative Likert-scale distribution points towards Scheherazade-IF doing very well in comparison to the Human upper bound. For the two scenarios we examined, Scheherazade-IF either meets Human performance or comes close to meeting human performance when data sparsity is controlled for. Our evaluation highlights the importance of data quality. We see an improvement in AI performance of 26.4% over random after the removal of just four plot graph nodes with sparse coverage. These lessons indicate that researchers should pay extra heed to consistent coverage in future generative systems informed by crowdsourcing or other machine learning techniques.

One of the shortcomings of our methodology was the inability to get data on remembered errors that we could trust as reliable. The remembered errors metric is designed to measure the extent to which commonsense errors are significantly noticed by players or glossed over. This is in contrast to recognized errors, in which players are asked to re-read and carefully analyze their game traces. The recognized errors metric provides ground-truth data on the accuracy of the AI system. An interesting future extension would be to manipulate participant perception of whether the plot graph was generated by human or AI authors.

Our human upper bound condition is limited by the fact that the human author had to work within the constraint of using plot events that matched those learned by the AI system and had to express typical understating of the given situations. This is especially the case in regards to the creativity within the interactive narrative stories. We do not doubt that a human could have come up with a significantly more creative experience. However, we still contend that a human could do no better than our human upper bound in working under the constraints of utilizing specific plot events.

6. CONCLUSIONS

In this paper we introduced the problem of open interactive narrative and give an overview of a crowdsourcing based approach, as implemented in the Scheherazade-IF system. As an open interactive narrative system, Scheherazade-IF learns everything that it needs to know to produce an interactive text-based experience in a just-in-time fashion. Evaluation of the Scheherazade-IF system shows that the system learns domain models for scenarios that are significantly improved over random models and, in some cases, approaching the performance of human-authored domain knowledge. Not surprisingly, Scheherazade-IF performance is partly a function of the quality of data provided by the crowdsourcing process. This stands as an important advancement not only in the field of interactive narrative, but in procedural game generation as well. The approach of crowdsourcing information to inform procedural systems holds promise towards solving problems that human individuals and computation cannot solve independently.

The current state of the Scheherazade-IF system points to a number of potentially impactful future directions. The current Scheherazade-IF system does not allow for particularly unusual experiences, as it generalizes towards an average understanding of stories. We wish to make the generated interactive narratives more creative. Crowdsourcing problems that could arise during the types of sociocultural events we currently look to, and how to resolve these problems, could serve as one means to inject additional creativity into the interactive narrative experiences. With additional conflict and resolutions, the inclusion of a drama manager into the system could further improve player experience by influencing the way in which an experience unfolds. Personalized drama management using data-driven player models of player preferences over possible experiences is theoretically compatible with Scheherazade-IF [23].

The stark difference in the movie date and robbery stories based on a slight sparseness of data indicates a further avenue for future work. Procedural game generation systems such as Scheherazade-IF that use just-in-time domain learning need to be able to assess their own models and determine when enough data is available and whether the data sufficiently covers all aspects of the game experience.

At this point, human authored interactive narrative still remains the most cost-effective means of generating an interactive narrative experience. However, *open interactive narrative* shows promise in reducing authorial burden in the near future. Scheherazade-IF and the lessons we learned in creating and evaluating it serve as a first step in creating human-quality interactive narrative with almost no human authoring required.

7. ACKNOWLEDGEMENTS

We gratefully acknowledge the NSF for supporting this research under NSF award 1350339. Special thanks to Alex Zook for advice on statistical analysis and draft comments.

8. REFERENCES

[1] Cook, M., Colton, S., Raad, A., and Gow, J. 2013. Mechanic miner: reflection-driven game mechanic discovery and level design. In *EvoGAMES*.

[2] Giannatos, S., Nelson, M. Cheong, Y., and Yannakakis, G. 2011. Suggesting new plot elements for an interactive story. *Proceedings of the 4th Workshop on Intelligent Narrative Technologies*, 25-30.

[3] Klein, D. and Manning, C. 2003. Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics,* 423-430.

[4] Li, B. 2014. Learning Knowledge to Support Domain-Independent Narrative Intelligence. Ph.D. Dissertation, Georgia Institute of Technology.

[5] Li, B., Lee-Urban, S., Johnston, G., and Riedl, M. 2013. Story generation with crowdsourced plot graphs. *Proceedings of the 27th AAAI Conference on Artificial Intelligence*.

[6] Li, B., Thakkar, M., Wang, Y., and Riedl, M. 2014. Storytelling with Adjustable Narrator Styles and Sentiments. *Proceedings of the 2014 International Conference on Interactive Digital Storytelling*, 1-12.

[7] Magerko, B. 2005. Evaluating preemptive story direction in the Interactive Drama Architecture. Journal of Game Development, 25-52.

[8] Miller, G. 1995. WordNet: A lexical database for English. Communications of the Association for Computing Machinery, 39–41.

[9] Nelson, M. and Mateas, M. 2005. Search-based drama management in the interactive fiction Anchorhead. *Proceedings of the 1st AAAI Conference on AI and Interactive Digital Entertainment*.

[10] Nelson, M. and Mateas, M. 2007. Towards Automated Game Design. *Proceedings of the 10th Congress of the Italian Association for Artificial Intelligence*.

[11] Nelson, M., Roberts, D., Isbell, Jr., C., and Mateas, M. 2006. Reinforcement learning for declarative optimization-based drama management. *Proceedings of the 5th International Joint Conference on Autonomous Agents & Multiagent Systems*.

[12] Orkin J. and Roy, D. 2009. Automatic learning and generation of social behavior from collective human gameplay. *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems*, 385-392.

[13] Porteus, J. and Cavazza, M. 2009. Controlling Narrative Generation with Planning Trajectories: The Role of

Constraints. *Proceedings of the 2nd International Conference on Interactive Digital Storytelling, 234-245.*

[14] Quinn, A. J. and Bederson, B. B. 2011. Human computation: a survey and taxonomy of a growing field. *Proceedings of The ACM SIGCHI Conference on Human Factors in Computing Systems*, 1403-1412.

[15] Riedl, M. and Bulitko, V. 2013. Interactive Narrative: An Intelligent Systems Approach. *AI Magazine*, 67-77.

[16] Riedl, M., Stern, A., Dini, D., and Alderman, J. 2008. Dynamic Experience Management in Virtual Worlds for Entertainment, Education, and Training. *International Transactions on System Science and Applications,* 23-42.

[17] Sharma, M., Mehta, M., Ontanon, S., and Ram, A. 2007. Player modeling evaluation for interactive fiction. *Proceedings of the 3rd AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment conference*

[18] Sina, S., Rosenfeld, A., and Kraus, S. 2014. Generating content for scenario-based serious games using crowdsourcing. *Proceedings of the 28th AAAI Conference on Artificial Intelligence*.

[19] Swanson, R. and Gordon, A. 2012. Say anything: Using Textual Case-Based Reasoning to Enable Open-Domain Interactive Storytelling. *ACM Transactions on Intelligent Interactive Systems.*

[20] Togelius, J. and Schmidhuber, J. 2008. An experiment in automatic game design. *Proceedings of the IEEE Symposium on Computational Intelligence and Games*, 15-18.

[21] Treanor, M., Schweizer, B., Bogost, I., and Mateas, M. 2012. The micro-rhetorics of Game-O-Matic. *Proceedings of the 7th International Conference on the Foundations of Digital Games*, 18-25.

[22] Weyhrauch, P. 1997. Guiding Interactive Drama, doctoral dissertation, tech. report CMU-CS- 97-109, School of Computer Science, Carnegie Mellon Univ.

[23] Yu, H. and Riedl, M. 2014. Personalized Interactive Narratives via Sequential Recommendation of Plot Points. *IEEE Transactions on Computational Intelligence and Artificial Intelligence in Games,* 174-182.

[24] Zook, A. and Riedl, M. 2014. Automatic Game Design via Mechanic Generation. *Proceedings of the 28th AAAI Conference on Artificial Intelligence*.