Constructing Social Networks from Unstructured Group Dialog in Virtual Worlds

Fahad Shah and Gita Sukthankar

Department of EECS
University of Central Florida
4000 Central Florida Blvd, Orlando, FL
sfahad@cs.ucf.edu, gitars@eecs.ucf.edu

Abstract. Virtual worlds and massively multi-player online games are rich sources of information about large-scale teams and groups, offering the tantalizing possibility of harvesting data about group formation, social networks, and network evolution. However these environments lack many of the cues that facilitate natural language processing in other conversational settings and different types of social media. Public chat data often features players who speak simultaneously, use jargon and emoticons, and only erratically adhere to conversational norms. In this paper, we present techniques for inferring the existence of social links from unstructured conversational data collected from groups of participants in the Second Life virtual world. We present an algorithm for addressing this problem, Shallow Semantic Temporal Overlap (SSTO), that combines temporal and language information to create directional links between participants, and a second approach that relies on temporal overlap alone to create undirected links between participants. Relying on temporal overlap is noisy, resulting in a low precision and networks with many extraneous links. In this paper, we demonstrate that we can ameliorate this problem by using network modularity optimization to perform community detection in the noisy networks and severing cross-community links. Although using the content of the communications still results in the best performance, community detection is effective as a noise reduction technique for eliminating the extra links created by temporal overlap alone.

Keywords: Network text analysis, Network modularity, Semantic dialog analysis

1 Introduction

Massively multi-player online games and virtual environments provide new outlets for human social interaction that are significantly different from both face-to-face interactions and non-physically-embodied social networking tools such as Facebook and Twitter. We aim to study group dynamics in these virtual worlds by collecting and analyzing public conversational patterns of Second Life users.

Second Life (SL) is a massively multi-player online environment that allows users to construct and inhabit their own 3D world. In Second Life, users control avatars, through which they are able to explore different environments and interact with other avatars in

a variety of ways. One of the most commonly used methods of interaction in Second Life is basic text chat. Users are able to chat with other users directly through private instant messages (IMs) or to broadcast chat messages to all avatars within a given radius of their avatar using a public chat channel. Second Life is a unique test bed for research studies, allowing scientists to study a broad range of human behaviors. Several studies on user interaction in virtual environments have been conducted in SL including studies on conversation [1] and virtual agents [2].

The physical environment in Second Life is laid out in a 2D arrangement, known as the SLGrid. The SLGrid is comprised of many regions, with each region hosted on its own server and offering a fully featured 3D environment shaped by the user population. The current number of SL users is estimated to be 16 million, with a weekly user login activity reported in the vicinity of 0.5 million [3].

Although Second Live provides us with rich opportunities to observe the public behavior of large groups of users, it is difficult to interpret who the users are communicating to and what they are trying to say from public chat data. Network text analysis systems such as Automap [4] that incorporate linguistic analysis techniques such as stemming, named-entity recognition, and n-gram identification are not effective on this data since many of the linguistic preprocessing steps are defeated by the slang and rapid topic shifts of the Second Life users. This is a hard problem even for human observers, and it was impossible for us to unambiguously identify the target for many of the utterances in our dataset. In this paper, we present an algorithm for addressing this problem, Shallow Semantic Temporal Overlap (SSTO), that combines temporal and language information to infer the existence of directional links between participants. One of the problems is that using temporal overlap as a cue for detecting links can produce extraneous links and low precision. To reduce these extraneous links, we propose the use of community detection. Optimizing network modularity reduces the number of extraneous links generated by overly generous temporal co-occurrence assumption but does not significantly improve the performance of SSTO.

There has been previous work on constructing social networks of MMORPG players, e.g., [5] looks at using concepts from social network analysis and data mining to identify tasks. Recent work [6] has compared the relative utility of different types of features at predicting friendship links in social networks; in this study we only examine conversational data and do not include information about other types of Second Life events (e.g., item exchanges) in our social networks.

2 Method

2.1 Dataset

We obtained conversation data from eight different regions in Second Life over fourteen days of data collection; the reader is referred to [7] for details. To study user dialogs, we examined daily and hourly data for five randomly selected days in the eight regions. In total, the dataset contains 523 hours of information over the five days (80,000 utterances) considered for the analysis across all regions. We did a hand-annotation of one hour of data from each of the regions to serve as a basis for comparison.

While there are corpora like [8], there has not been any body of work with an online chat corpus in a multi-user, open-ended setting — the characteristics of this dataset. In such situations it is imperative to identify conversational connections before proceeding to higher level analysis like topic modeling, which is itself a challenging problem. We considered several approaches to analyzing our dialog dataset, ranging from statistical NLP approaches using classifiers to corpus-based approaches using tagger/parsers; however we discovered that there is no corpus available for group-based online chat in an open-ended dialog setting. It is challenging even for human labelers to annotate the conversations themselves due to the large size of the dataset and the ambiguity in a multi-user open-ended setting. Furthermore, the variability of the utterances and the nuances such as emoticons, abbreviations and the presence of emphasizers in spellings (e.g., "Yayyy") makes it difficult to train appropriate classifiers. Parser/tagger-based approaches perform poorly due to the lack of corpus and inclusion of non-English vocabulary words.

Consequently, we decided to investigate approaches that utilize non-linguistic cues such as temporal co-occurrence. Although temporal co-occurrence can create a large number of false links, many aspects of the network group structure are preserved. Hence we opted to implement a two-pass approach: 1) create a noisy network based solely on temporal co-occurrence, 2) perform modularity detection on the network to detect communities of users, and 3) attempt to filter extraneous links using the results of the community detection.

2.2 Modularity Optimization

In prior work, community membership has been successfully used to identify latent dimensions in social networks [9] using techniques such as eigenvector-based modularity optimization [10] which allows for both complete and partial memberships. As described in [10], modularity (denoted by Q below) measures the chances of seeing a node in the network versus its occurrence being completely random; it can be defined as the sum of the random chance $A_{ij} - \frac{k_i k_j}{2m}$ summed over all pairs of vertices i, j that fall in the same group, where s_i equals 1 if the two vertices fall in the same group and -1 otherwise:

$$Q = \frac{1}{4m} \sum (A_{ij} - \frac{k_i k_j}{2m}) s_i s_j. \quad (1)$$

If B is defined as the modularity matrix given by $A_{ij} - \frac{k_i k_j}{2m}$, which is a real symmetric matrix and s column vectors whose elements are s_i then Equation 1 can be written as $Q = \frac{1}{4m} \sum_{i=1}^{n} (u_I^T s)^2 \beta_i$, where β_i is the eigenvalue of B corresponding to the eigenvector u (u_i are the normalized eigenvectors of B so that $s = \sum_i^n a_i u_i$ and $a_i = u_i^T s$). We use the leading eigenvector approach to spectral optimization of modularity as described in [11] for the strict community partitioning (s being 1 or -1 and not continuous). For the maximum positive eigenvalue we set $s = 1$ for the corresponding element of the eigenvector if its positive and negative otherwise. Finally we repeatedly partition a group of size n_g in two and calculate the change in modularity measure given by $\Delta q = \frac{1}{4m} \sum_{i,j \epsilon g} [B_{ij} - \delta_{ij} \sum_{k \epsilon g} B_{ik}] s_i s_j$, where δ_{ij} is the Kronecker δ symbol, terminating if the change is not positive and otherwise choosing the sign of s (the partition) in the same way as described earlier.

2.3 Shallow Semantics and Temporal Overlap Algorithm (SSTO)

Because of an inability to use statistical machine learning approaches due to the lack of sufficiently labeled data and absence of a tagger/parser that can interpret chat dialog data, we developed a rule-based algorithm that relies on shallow semantic analysis of linguistic cues that commonly occur in chat data including mentions of named entities as well as the temporal co-occurrence of utterances to generate a to/from labeling for the chat dialogs with directed links between users. Our algorithm employs the following types of rules:

salutations: Salutations are frequent and can be identified using keywords such as "hi", "hello", "hey". The initial speaker is marked as the *from* user and users that respond within a designated temporal window are labeled as *to* users.

questions: Question words (e.g., "who", "what", "how") are treated in the same way as salutations. We apply the same logic to requests for help (which are often marked by words such as "can", "would").

usernames: When a dialog begins or ends with all or part of a username (observed during the analysis period), the username is marked as *to*, and the speaker marked as *from*.

second person pronouns: If the dialog begins with a second person pronoun (i.e., "you", "your"), then the previous speaker is considered as the *from* user and the current speaker the *to* user; explicit mentions of a username override this.

temporal co-occurrences: Our system includes rules for linking users based on temporal co-occurrence of utterances. These rules are triggered by a running conversation of 8–12 utterances.

This straightforward algorithm is able to capture sufficient information from the dialogs and is comparable in performance to SSTO with community information, as discussed below.

2.4 Temporal Overlap Algorithm

The temporal overlap algorithm consists of using the temporal co-occurrence to construct the links. It exploits the default timeout in Second Life (20 minutes) and performs a lookup for 20 minutes beginning from the occurrence of a given username and constructs an undirected link between the speakers and this user. This process is repeated for all users within that time window (one hour or day) in 20 minute periods. This algorithm gives a candidate pool of initial links between the users without considering any semantic information. Later, we show that incorporating community information from any source (similar time overlap or SSTO based) and on any scale (daily or hourly) enables us to effectively prune links, showing the efficacy of mining community membership information.

2.5 Incorporating Community Membership

Our dataset consists of 5 randomly-chosen days of data logs. We separate the daily logs into hourly partitions, based on the belief that an hour is a reasonable duration for social interactions in a virtual world. The hourly partitioned data for each day is used to

generate user graph adjacency matrices using the two algorithms described earlier. The adjacency matrix is then used to generate the spectral partitions for the communities in the graph, which are then used to back annotate the tables containing the to/from labeling (in the case of the SSTO algorithm). These annotations serve as an additional cue capturing community membership. Not all the matrices are decomposable into smaller communities so we treat such graphs of users as a single community.

There are several options for using the community information — we can use the community information on an hourly- or daily basis, using the initial run from either SSTO or the temporal overlap algorithms. The daily data is a long-term view that focuses on the stable network of users while the hourly labeling is a fine-grained view that can enable the study of how the social communities evolve over time. The SSTO algorithm gives us a conservative set of directed links between users while the temporal overlap algorithm provides a more inclusive hypothesis of users connected by undirected links.

For the SSTO algorithm, we consider several variants of using the community information:

SSTO: Raw SSTO without community information;
SSTO+LC: SSTO (with loose community information) relies on community information from the previous run only when we fail to make a link using language cues.
SSTO+SC: SSTO (with strict community information) always uses language cues in conjunction with the community information.

For the temporal overlap algorithms, we use the community information from the previous run.

TO: Raw temporal overlap algorithm without community information;
TO+DT Temporal overlap plus daily community information;
TO+HT Temporal overlap plus hourly community information.

3 Results

In this section we summarize the results from a comparison of the social networks constructed from the different algorithms. While comparing networks for similarity is a difficult problem [12], we restrict our attention to comparing networks as a whole in terms of the link difference (using Frobenius norm) and a one-to-one comparison for the *to* and *from* labelings for each dialog on the ground-truthed subset (using precision and recall).

3.1 Network Comparison Using the Frobenius Norm

We constructed a gold-standard subset of the data by hand-annotating the to/from fields for a randomly-selected hour from each of the Second Life regions. It is to be noted that there were instances where even a human was unable to determine the person addressed due to the complex overlapping nature of the dialogs in group conversation in an open ended setting (Table 2).

To compare the generated networks against this baseline, we use two approaches. First we compute a Frobenius norm [13] for the adjacency matrices from the corresponding networks. The Frobenius norm is the matrix norm of an $M \times N$ matrix A and is defined as:

$$\|A\| = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}|a_{ij}|^2}. \quad (2)$$

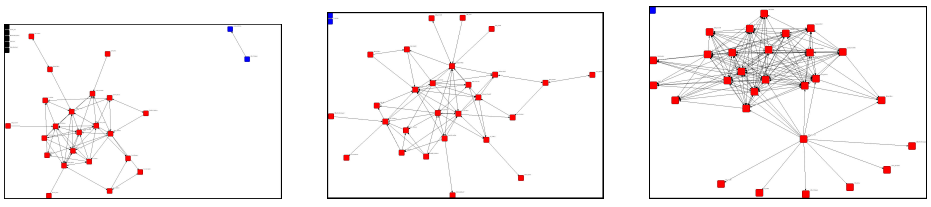
The Frobenius norm directly measures whether the two networks have the same links and can be used since the networks consists of the same nodes (users). Thus, the norm serves as a measure of error (a perfect match would result in a norm of 0). Table 1 shows the results from this analysis.

Table 1. Frobenius norm: comparison against hand-annotated subset.

	SSTO	SSTO+LC	SSTO+SC	TO	TO+DT	TO+HT
Help Island Public	35.60	41.19	46.22	224.87	162.00	130.08
Help People Island	62.23	60.50	66.34	20.29	20.29	54.88
Mauve	48.45	45.11	51.91	58.44	58.44	49.89
Morris	24.67	18.92	20.76	43.12	37.54	38.98
Kuula	32.12	30.75	32.66	83.22	73.15	77.82
Pondi Beach	20.63	21.77	21.56	75.07	62.62	71.02
Moose Beach	17.08	18.30	21.07	67.05	53.64	50.97
Rezz Me	36.70	39.74	45.78	38.72	39.01	41.10
Total error	277.48	**276.28**	306.30	610.78	507.21	514.74

3.2 Direct Label Comparisons

The second quantitative measure we present is the head-to-head comparison of the to/from labelings for the dialogs using any of the approaches described above (for SSTO) against the hand annotated dialogs. This gives us the true positives and false positives for the approaches and allows us to see which one is performing better on the dataset, and if there is an effect in different Second Life regions. Table 2 shows the results from this analysis.



(a) Hand labeled network. (b) SSTO labeled network. (c) TO labeled network.

Fig. 1. Networks from different algorithms for one hour in the Help Island Public region.

Table 2. Precision/Recall values for one-to-one labeling comparison.

		Help Island Public	Help People Island	Mauve	Morris	Kuula	Pondi Beach	Moose Beach	Rezz Me
Total Dialogs		360	184	128	179	227	144	128	97
Hand Labeled	recall	0.6278	0.9076	0.9453	0.6983	0.8370	0.6944	0.6797	0.8866
	total	226	167	121	125	190	100	87	86
SSTO+SC	match	61	59	49	43	63	27	12	23
	precision	0.2607	0.6629	0.6364	0.4216	0.4632	0.3971	0.2105	0.4600
	recall	0.2699	0.3533	0.4050	0.3440	0.3316	0.2700	0.1379	0.2674
	F-Score	0.2652	0.4609	0.4204	0.3789	0.3865	0.3214	0.1667	0.3382
	total	234	89	77	102	136	68	57	50
SSTO+LC	match	61	51	37	39	52	26	12	15
	precision	0.3005	0.6456	0.6607	0.4643	0.4561	0.4194	0.2667	0.4688
	recall	0.2699	0.3054	0.3058	0.3120	0.2737	0.2600	0.1379	0.1744
	F-Score	0.2844	0.4146	0.4181	0.3732	0.3421	0.3210	0.1818	0.2542
	total	203	79	56	84	114	62	45	32
SSTO	match	76	68	51	45	66	30	20	27
	precision	0.3065	0.7083	0.6145	0.4500	0.4748	0.4225	0.3077	0.4576
	recall	0.3363	0.4072	0.4215	0.3600	0.3474	0.3000	0.2299	0.3140
	F-Score	0.3207	0.5171	0.5000	0.3617	0.4012	0.3509	0.2299	0.3724
	total	248	96	83	100	139	71	65	59

4 Conclusion

For the temporal overlap algorithm (TO), the addition of the community information reduces the link noise, irrespective of the scale — be it hourly or daily. This is shown by the decreasing value of the Frobenius norm in all the cases as compared to the value obtained using temporal overlap algorithm alone. In general shallow semantic approach (SSTO) performs the best and is only improved slightly by the loose incorporation of community information. For the SSTO algorithm, the daily or hourly community partition also does not affect the improvement. Table 2 shows how the dialog labeling generated from various algorithms agrees with the ground truth notations produced by a human labeler. Since TO only produces undirected links, we do not include it in the comparison. Plain SSTO generally results in a better precision and recall than SSTO plus either strict or loose community labeling. This is further confirmed from the visualizations for one hour of data in a day (both chosen randomly from the dataset) for each of the ground-truth, SSTO and TO as shown in figure 1, where the network from SSTO more closely resembles the ground-truth as compared to the on from TO.

The challenging nature of this dataset is evident in the overall low precision and recall scores, not only for the proposed algorithms but also for human labelers. We attribute this largely to the inherent ambiguity in the observed utterances. Among the techniques, SSTO performs best, confirming that leveraging semantics is more useful than merely observing temporal co occurrence. We observe that community informa-

tion is not reliably informative for SSTO but does help TO, showing that link pruning through network structure is useful in the absence of semantic information.

Acknowledgments

This research was funded by AFOSR YIP award FA9550-09-1-0525.

References

1. Weitnauer, E., Thomas, N.M., Rabe, F., Kopp, S.: Intelligent agents living in social virtual environments bringing Max into Second Life. In: International Working Conference on Intelligent Virtual Agents. (2008)
2. Bogdanovych, A., Simoff, S., Esteva, M.: Virtual institutions: Normative environments facilitating imitation learning in virtual agents. In: International Working Conference on Intelligent Virtual Agents. (2008)
3. Second Life: Second Life Economic Statistics (2009) Retrieved July 2009 http://secondlife.com/whatis/economy_stats.php.
4. Carley, K., Columbus, D., DeReno, M., Bigrigg, M., Diesner, J., Kunkel., F.: Automap users guide 2009. Technical Report CMU-ISR-09-114, Carnegie Mellon University, School of Computer Science, Institute for Software Research (2009)
5. Shi, L., Huang, W.: Apply social network analysis and data mining to dynamic task synthesis to persistent MMORPG virtual world. In: Proceedings of Intelligent Virtual Agents. (2004)
6. Kahanda, I., Neville, J.: Using transactional information to predict link strength in online social networks. In: Proceedings of the Third International Conference on Weblogs and Social Media. (2009)
7. Shah, F., Usher, C., Sukthankar, G.: Modeling group dynamics in virtual worlds. In: Proceedings of the Fourth International Conference on Weblogs and Social Media. (2010)
8. Mann, W.C.: The dialogue diversity corpus (2003) Retrieved Nov 2010 http://www-bcf.usc.edu/~billmann/diversity/DDivers-site.htm.
9. Tang, L., Liu, H.: Relational learning via latent social dimensions, in 'kdd '09. In: Proceedings di of the 15th ACM SIGKDD international ti conference on Knowledge, ACM (2009) 817–826
10. Newman, M.: Modularity and community structure in networks. In: Proceedings of the National Academy of Sciences. Volume 103. (2006) 8577–8582
11. Newman, M.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E 74, 036104 (2006)
12. Prulj, N.: Biological network comparison using graphlet degree distribution. Bioinformatics **23**(2) (2007)
13. Golub, G.H., Loan, C.F.V.: Matrix Computations. 3rd edn. JHU Press (1996)