VOL. 10, NO. 17, SEPTEMBER 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

FRIEND MATCHING USING PROBABILISTIC TOPIC MODEL

Vidya R. and Nishada S. G.
Department of Computer Science and Engineering, Mohandas College of Engineering, Kerala, India

ABSTRACT

Recommender Systems provide suggestions for users to guide in various decision-making processes.The recommender systems can be defined by the purpose of recommendation, mechanism and data gathering. Recommendation system for social networks are different since the item recommended are rational human beings. The paper focuses on designing a friend matching system by analyzing user lifestyles as common criteria. Large amount of data collected from various users create high dimensional data. In order to resolve this, probabilistic topic modeling is used. Content based machine learning approaches are used to find out suspicious users in the recommendation system. The results are evaluated based on the datasets created from the real world users.

Keywords: friend-of-friend, probabilistic topic modeling, latent dirichlet allocation.

INTRODUCTION

Online Social network services have thrived in great popularity recent years. Social networking sites have attracted tremendous numbers of users and play an important role in online interaction. By connecting users with similar common interests, online social networks open up a new channel for information sharing and social networking. The open ended nature of their applications motivates rich user-generated content, including tags, text document, multimedia, and so on.

One fundamental phenomenon in social network services is friendship formation. Members make friends with each other through social interactions and information exchange. A member in online social network may be frustrated to find new friends from a tremendous number of irrelevant users. Suggesting relevant users with common interests to each individual can help improve user experience. Most social network websites match members based on the number of mutual friends. This method suffers the drawback of interest mismatch and it is useless to expand the circle of the members. According to these studies the rules to group people together include: 1) habits or life style; 2) attitudes; 3) tastes; 4) moral standards5) economic level and 6) people they already know. Apparently, rule #3 and rule #6 are the mainstream factors considered by existing recommendation systems. Grouping friends by means of their habits were not adopted widely. This is due to the difficulty in analyzing user habits in real time. Rather, life styles are usually closely correlated with daily routines and activities. Therefore, if it is possible to obtain the daily routines of users then we can recommend new friends based on the common habits.

The remainder of this paper is organized as follows:

Section II provides an overview on related works in friend recommendation system. Section III provides system description of the proposed work. It explains the model for capturing user habits, how to identify lifestyles and generate lifestyle vector. This also explains an efficient friend recommendation algorithm. Section IV deals with datasets used in the work. Section V presents the results of applying the proposed method and Section VI concludes.

RELATED WORK

Xiao Yu, Ang Pan, Lu-An Tang [1] proposed a friend recommendation mechanism by identifying geographically related friends. This friend recommendation approach considered the users current geography. Similarity among user interests were not included which lacks the user's preference on friend selection in real world. Alvin Chin, Bin Xu, Hao Wang [2] proposed a friend recommendation based on physical context. Here physical context is based on meetings and encounters. The method uses the intuition that people who meet for a conference can be recommended as friends. Jeff Naruchitparames, Mehmet Hadi Gunes and Sushil J. Louis [3] proposed a friend recommendation based on network topology and genetic algorithm. This approach also ensures the likelihood of a person pursing a friendship of someone they know than someone they do not know. Zhing Wang [4] proposed Friendbook which extracts user habits based on sensors like accelerometer and gyroscope. The user activities captured were limited to indoor activities. The proposed method captures user habits with the help of smartphones. Inspired by advancements on probabilistic topic modeling; the study proposes a generative model based on LDA (Latent Dirichlet Allocation) for dimensionality reduction. Thus topic model can generate relevant interests of a user .The proposed method provides a modified friend recommendation algorithm that enhances accuracy compared to traditional algorithms.

THE PROPOSED SCHEME

Friend Matcher is an android application that generates new friendship recommendations to the users. Sensors in Smartphone are used to capture user's daily habits. The user habits are captured and analyzed by means of probabilistic topic modeling. The habits captured can be divided based on regular activities and user behavior. Regular activities comprises of lifestyles like traveler, teacher, student and so on. The parameters we

ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

consider for lifestyle analysis are SMS, search history, application installed and detection of suspicious activity. The previous works considered various indoor sensor readings to detect user habits. The flowchart below depicts the model used for capturing user life habits.

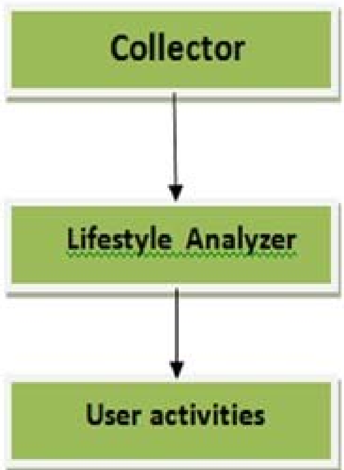

Figure-1. The flowchart.

The daily routines of the user are captured by the collector module. The data captured is fed to the lifestyle analyzer. The lifestyle analyzer uses topic modeling to extract meaningful life habits of the user. The system provides five modules namely data collection layer, lifestyle extraction, similarity calculation, graph creation and integrating feedback to the system.

Data Collection

Interest of the user is analyzed by location based services, search history, SMS and applications installed in the smartphone. The system updates the server about GPS coordinates and the services accessed by the user. Further by the application of text mining interests can be captured from SMS and search history. For this an unsupervised learning approach called probabilistic topic modeling is used. The same word can appear in different topics due to its ambiguity.

Lifestyle Analyzer

Topic modeling uses a non deterministic algorithm called as LDA [11]. This will identify lifestyles of the user. In LDA [11], features of each document can be divided into two parts: features relevant to user lifestyles and irrelevant features like stopwords.LDA outputs set of meaningful words and corresponding ranks. The algorithms for user lifestyle generation are as follows:

Table-1. Identify user lifestyles using topic model.

Step 1: For each label
For each user document in the collector
Train the LDA model for k topics
Obtain top M words for topic K by posterior probability

Step 2: Merge all top M words for all documents
Generate the lifestyle vector.

Each user's collector module consists of several lifestyles. System represents each user in terms of a lifestyle vector. Lifestyle vectors are represented as probabilities of several lifestyles over collector. Collector module logs all user routines. Let the lifestyles of user1 be lifestyle1, lifestyle2. Then we write:

L1= [p(lifestyle1|collector1), p(lifestyle2|collector1) (1)

The topics or the lifestyles extracted are updated in the database. Based on the parameters we consider for lifestyle analysis, the equation for lifestyle vector of a user is written as:

$$LF_{U_i} = LF_{GPS} + LF_{URL} + LF_{SMS} + LF_{APP} + LF_{SP} \quad (2)$$

Thus lifestyle vector of user depends on values obtained from GPS, visited URLs, messages, applications in the mobile and detection of suspicious activity. The similarity between users is calculated based on the cosine similarity. A threshold value is set and only those values above the threshold are considered for friendship calculation. Based on the values obtained, a graph is plotted with users as nodes and similarity value as edges. For friendship recommendation, a friend up to level3 is considered from the friendship graph. This is to ensure that recommendations are not obtained from complete strangers.

In order to check the suspicious activity a spam classifier is used. The system is trained with set of spam and non-spam messages. Based on this training set, system can identify whether a lifestyle belongs to spam category or not. A suspicious user thus identified is not eliminated from the system. Instead a negative score is assigned to such a user. Figure-2 shows the modules used in the system.

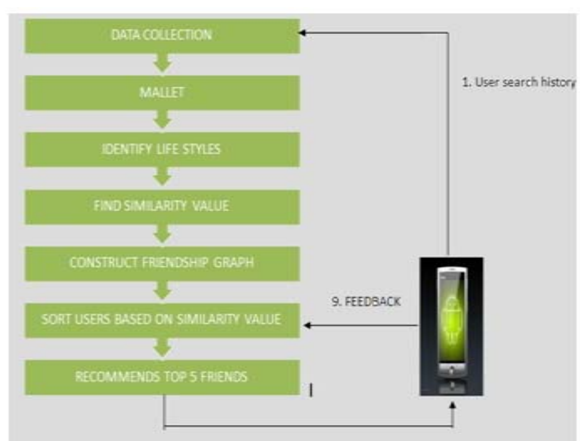

Figure-2. Modules in the system.

VOL. 10, NO. 17, SEPTEMBER 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

The similarity value calculated between the users provides the friendship score. These score values are sorted in the descending order. Corresponding users with the top five scores are recommended as new friends. The users can express the satisfaction on received recommendations using the feedback module. An algorithm for friend matching is proposed which will select top five friends based on the recommendation score. Modules used in the system are:

Similarity Calculation

In order to find interest similarity between users we use cosine similarity as the metric. Let L1 := [p(z1|d1),p(z2|d1),p(z3|d1)....p(zt|d1];L1:=[p(z1|d2),p(z2|d2),p(z3|d2)....p(zt|d2]where z represents the user lifestyles and d represents the document in collector.
Therefore, the similarity of habits between user1 and user2 is denoted by:

Sim (u1, u2):= cos (L1, L2)

Algorithm1: Find Similarity between users

Step 1: **For** each life style z_k, the probability is not zero
If $z_{k=}$ sms then Check suspicion (); **End if**

SIM_{score} =lvector (i)*lvector (j)
k=k+1
Endfor
Step 2: **For** each user with no match in lifestyle **do**
Initialize the similarity value, S (i, j) =0
Endfor

Friend Graph Creation

Using similarity metric, we find the interest similarity between users. Then a friend matching graph is created. It is a weighted graph G=(V,E,W)where users in the system are represented as vertices and the similarity score as edges .Only if there is similarity in lifestyles an edge will be created. The similarity score obtained should be greater than the threshold. Such edges are only inserted. Thus weight of an edge W (i,j):= S(i,j). Here we use friend of friend (FoF) approach so as to avoid receiving recommendations from complete strangers. For friendship recommendation, friends up to level3 are considered.

Friend Suggestion

A user when submits a query for friend suggestion, pool creation and rank calculation takes place .For each user with similar interests the system calculates friendship score. Suppose the system finds 'j' similar users for user1.Then, rank calculation is given by the equation Fscore (u1, u2):

$$Sim_{gps} + Sim_{mp3} + Sim_{app} + Sim_{sms} \quad (3)$$

The system first creates a pool of users. Among the users ,new friends will be suggested .This pool of users are created from the friend graph where the system starts capturing details of friends of friends and navigate deep down the tree till level 3.We want to limit recommendations from strangers, Hence we set the threshold as 3.Algorithm is illustrated below:

Algorithm2: Create list for recommendation

Step 1: alreadyfriends := getFriends ();
alreadyfriends_level:= 0
temp ← alreadyfriends
Step 2: **while** (temp≠null)
For each user 'x'from temp
current← temp (0) //first entry in temp;
remove.temp (0);
Step 3: **if** current. Level < 3 then
Frndlist:= getfrndsof_current();
For each user 'y' in the frndlist
y.level:= current. Level + 1
temp.add (y.)
recomm_list:=y .
End for
End if end for
End while

The system selects friends for recommendation from the pool created. The friend suggestion method is explained with the help of algorithm.

Algorithm3: Friend Suggestion Algorithm

Step 1 Extract user Ui, lifestyle vector using topic modeling.
Step 2: Compute lifestyle vector for each
$LV_{U_i} = \{LF_{GPS}, LF_{SMS}, LF_{app}, LF_{mp3}\}$
Step 3: get recommend_ list ()
Step 4: **For** each user in recommend _list,
Getsimilarity ()
Draw the friendship graph();
$Fscore: Sim_{gps} + Sim_{mp3} + Sim_{app} + Sim_{sms}$
Endfor
Step 5: sort all users in decreasing order according to score
Step 6: Output the top five users from the sorted list

Identify Suspicious Activity

In order to check the suspicious activity a spam classifier is used. Using topic modeling, the user SMS is used for monitoring this activity. Initially the system is trained with set of spam and non-spam messages. Based on this training set, system can identify whether a lifestyle belongs to spam category or not. We find out the score for user sms based on the rank calculation. Rank is expressed in terms of probability distribution. Difference between the ranks is calculated to identify if the user is suspicious. A suspicious user thus identified is not eliminated from the system. Instead a negative score is assigned to such a user. The server maintains a list of suspicious users.

VOL. 10, NO. 17, SEPTEMBER 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

Algorithm 4: Find malicious users

Check suspicion ()
Step1: Train the system with set of spam and non spam messages
Step 2: **For** each user in the friend list
Check the distribution of spam and non spam
Score=rank (no spam) - rank (spam)
Step 3: **If** (score <0) **then**
Raise suspicion
$LF_{sms} = -- (LF_{F_i})$
End if End for

EXPERIMENTAL DATA

We rely on datasets directly collected from the user's Smartphone. Based on the GPS coordinates, the services accessed by the users are obtained. Similarly the user browsed URLs are captured from the mobile device in real time. The SMS database used in the study contains one set of SMS messages in English of 5,574 messages, tagged according being 'not spam' and 'spam'. Table-3 provides some statistics of SMS dataset.

Table-2. SMS dataset details.

	No spam	spam
Total SMSes	4827	747
Average number of words	0.133	0.167
Average Presence of urls	0	2

Following are the observations obtained from analyzing SMS. Numbers of special characters were high for no spam category as the messages were informal. Presence of URL was found more in spam category. Figures below represent tag cloud generated from analyzing SMS messages using topic modeling. The system is trained with set of spam and non spam messages together. Suspicion is raised whenever data collected from the user contains threat beyond a threshold value.

Figure-3. Nonspam and spam words identified by topic modelling.

EXPERIMENTAL RESULTS

The results were evaluated with the help of eight Smartphone users. Each user used a Smartphone with Friend Matcher installed. The lifestyles considered were office, student, researcher, lecturer and business. Using topic modeling lifestyles of the users were classified. Classification result depends on the value set as number of topics .It is observed that as the number of topics is set to 100 ,classification results are better Figure-4 shows the lifestyle distribution for a particular user.

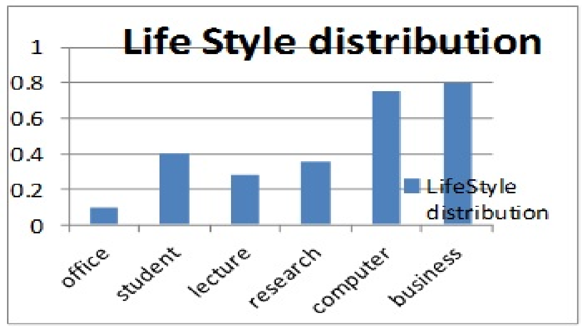

Figure-4. Lifestyle distribution.

The lifestyle distribution of each user can be graphically represented based on the probability distribution. For evaluation purpose six lifestyles are considered. The bar chart indicates the relevance of each lifestyle to a user. Thus the lifestyle business contributes to 0.8, computers to 0.7and student to 0.4. Thus lifestyle vector is denoted by L1={0.8,0.7,0.4…}.The final recommendation scores obtained is shown below.

UserID	UserName	Weight
12	Gurbachan	2.098083496093...
72	Arijit	2.098083496093...
87	Debashish	2.288818359375...
88	Fanindra	2.288818359375...
117	Balaji	1.907348632812...
25	Amal	9.536743164062...
85	Akshat	9.536743164062...

Figure-5. Recommendation scores.

The interesting factor in our work is the consideration of more realistic parameters in the system. The parameters considered depend on the lifestyle values. The lifestyle values affect the friendship scores. This in turn influences the efficiency of friend recommendation system. Here we plot two graphs with user size as x-axis and lifestyle weights as y-axis. The Friend Graph is plotted with lifestyle weights obtained by considering parameters like location based services and visited URLs .In this case lifestyle weight is obtained by:

$$LF_{U_i} = LF_{GPS} + LF_{URL} \tag{4}$$

S&P Friend graph considers the modified parameters like SMS, applications installed, music and suspicion detection also. The modified parameters change the lifestyle weights. Thus we write:

$$LF_{U_i} = LF_{GPS} + LF_{URL} + LF_{SMS} + LF_{APP} + LF_{SP} \tag{5}$$

We observe difference in lifestyle weights between the two graphs due to the modified parameters considered. This difference in lifestyle weight is the error we have identified using our proposed method. This justifies that our work enhances accuracy in recommending friends.

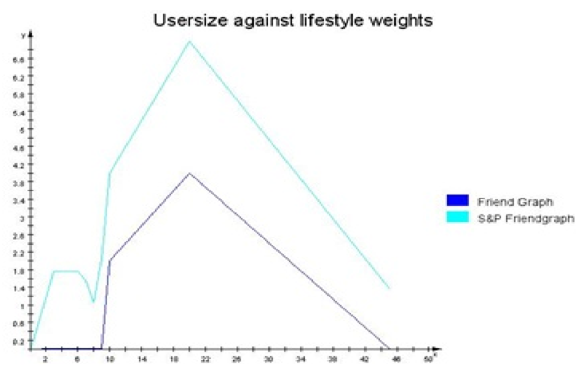

Figure-6: Observed lifestyle weights

CONCLUSIONS

The work proposes an approach quite different from the existing friend recommendation mechanisms. The latter relies on collaborative method of recommending friends. Whereas the proposed work extracts user habits by tracking daily activities of the user. It recommends potential friends to users if they share similar life styles. The result shows that the recommendations accurately reflect the preferences of users in choosing friends.

As the system extracts the user interests, this can be used to recommend several interesting features to a particular user. Thus the user receives meaningful information according to his tastes. The friendship parameters used in the system helps to increase user's credibility in system. As a means to offer privacy, the complete information of the user is not provided while recommendation. System just displays the relevant score calculated.

ACKNOWLEDGEMENTS

I would like to thank all the participants who participated for data collection. I extend my gratitude to Prof. Nishada S.G for providing guidance to this work.

REFERENCES

[1] Xiao Yu, Ang Pan, Lu-An Tang," Geo-Friends Recommendation in GPS-based Cyber-physical Social Network", 2011, pp.361-368.

[2] Alvin Chin, Bin Xu, Hao Wang," Who should I add as a "friend"? study of friend recommendations using proximity and homophily",MSM,2013,pp.7.

[3] J. Naruchitparames, M.H. Gunes, and S.J. Louis, "Friend recommendations in social networks using genetic algorithms and network topology";in Proc. IEEE Congress on Evolutionary Computation, 2011, pp.2207-2214.

[4] J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford. A Hybrid Discriminative/Generative Approach for Modeling Human Activities. Proc. of IJCAI, pages 766-772, 2005.

[5] N. Eagle and A. S. Pentland. Reality Mining: Sensing Complex Co-social Systems. Personal Ubiquitous Computing, 10(4):255-268, March 2006.

[6] K. Farrahi and D. Gatica-Perez. Probabilistic mining of socio-geographic routines from mobile phone data. Selected Topics in Signal Processing, IEEE Journal of, 4(4):746-755, 2010.

[7] W. H. Hsu, A. King, M. Paradesi, T. Pydimarri, and T. Weninger. Collaborative and structural recommendation of friends using weblog-based social network analysis. Proc. of AAAI Spring Symposium Series, 2006.

[8] L. Gou, F. You, J. Guo, L. Wu, and X. L. Zhang. Sfviz: Interest based friends' exploration and recommendation in social networks. Proc. of VINCI, page 15, 2011.

[9] T. Huynh, M. Fritz, and B. Schiel. Discovery of Activity Patterns using Topic Models. Proc. of UbiComp, 2008.

[10] ZhingWang,long Liao. Friendbook: A semantic friend recommendation method system for social networks Mobile Computing, IEEE Transactions on Vol4, pages538-551, 2014.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993-1022, 2003.

[12] Q. Li, J. A. Stankovic, M. A. Hanson, A. T. Barth, J. Lach, and G. Zhou. Accurate, Fast Fall Detection Using Gyroscopes and Accelerometer-Derived Posture Information. Proc. of BSN, pages 138-143, exploration and recommendation in social networks. Proc. of

VINCI, page 15, 2011.

[13] Hu. Y. Jun, Sun Yat-sen"Chinese Short Text Classification Based on Topic Model with High frequency feature Expansion," Journal of Multimedia, Vol. 8, No. 4, August 2013.

[14] Farrahi and D. Gatica-Perez," Discovering Routines from Large scale Human Locations using Probabilistic Topic Models, "ACM Transactions on Intelligent Systems and Technology (TIST), 2(1), 2011.

[15] D. Nagamalai, E. Renault and M. Dhanushkodi, "Trust enhanced recommendation of friends in social network using genetic algorithm to learn user preference".

[16] Trends in Computer Science, Engineering and Information Technology Communications in Computer and Information Science Volume 204, 2011, pp 476-485.

[17] Z.Wang,C.E,Taylor,H.Qui.Demo:Friendbook Privacy preserving friend matching based on shared interests ;in Proc of ACM,pages 397-398, 2011.

[18] Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma. Understanding Transportation Modes Based on GPS Data for Web Applications. ACM Transactions on the Web (TWEB), 4(1):1–36, 2010.

[19] L. Bian and H. Holtzman. Online friend recommendation through personality matching and collaborative filtering. In Proc. of UBICOMM, pages 230–235, 2011.

[20] Kumar, R., Novak, J., and Tomkins, A. Structure and Evolution of Online Social Networks. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. (Philadelphia, PA, USA, August 20-23, 2006). KDD'06. ACM Press, New York, NY, 611-617.

[21] Gediminas Adomavicius , Alexander Tuzhilin, Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, IEEE Transactions on Knowledge and Data Engineering, v.17 n.6, p.734-749, June 2005.

[22] Basu, C., Hirsh, H., and Cohen W. 2001. Recommendation as classification: Using social and content-based information in recommendation. Recommender systems papers from 1998 workshop, Tech. rep. WS-98-08, AAAI Press.

[23] Toine Bogers , Antal van den Bosch, Comparing and evaluating information retrieval algorithms for news recommendation, Proceedings of the 2007 ACM conference on Recommender systems, October 19-20, Minneapolis, MN, USA

[24] Brunato, M., Battiti, R., Villani, A., and Delal, A. 2002. A location dependent recommender system for the Web. Tech. rep. DIT-02-093, University of Trento

[25] Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., and Sartin, M. 1999. Combining content-based and collaborative fitters in an online newspaper. In Proceedings of the ACM SIGIR Workshop on Recommender Systems: Algorithms and Evaluation.

[26] Q. Li, J. A. Stankovic, M. A. Hanson, A. T. Barth, J. Lach, and G. Zhou. Accurate, Fast Fall Detection Using Gyroscopes and Accelerometer-Derived Posture Information. Proc. of BSN, pages 138-143, exploration and recommendation in social networks. Proc. of VINCI, page 15, 2011.

[27] Wan-Shiou Yang , Hung-Chi Cheng , Jia-Ben Dia, A location-aware recommender system for mobile shopping environments, Expert Systems with Applications: An International Journal, v.34 n.1, p.437-445, January, 2008

[28] Yu Zheng , Yukun Chen , Xing Xie , Wei-Ying Ma, GeoLife2.0: A Location-Based Social Networking Service, Proceedings of the 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware, p.357-358, May 18-20, 2009.

[29] Zheng, Y., Xie, X., and Ma, W. Y. 2010 d. GeoLife: A collaborative social networking service among user, location and trajectory. IEEE Data Engin. Bull. 33, 2, 32--40.[29]topology", In Proc. IEEE Congress on Evolutionary Computation, pp.2207-2214.

[30] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. Application of dimensionality reduction recommender system: A case study, in Proceedings of the ACM WebKDD Workshop.

[31] Chen, J., Geyer, W., Dugan, C., and Guy, I. Make new friends, but keep the old: recommending people on social networking sites. Proc. CHI, pp. 201-210. 2009.

[32] Kumar, R., Novak, J., and Tomkins, A. Structure and Evolution of Online Social Networks. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. (Philadelphia, PA, USA, August 20-23, 2006). KDD'06. ACM Press, New York, NY, 611-617.