# Application of Novel Features in Complex Network for Analyzing Virtual Community

## Zhen Zhang<sup>1</sup>, Qingchun Meng<sup>2</sup>, Xiaoxia Rong<sup>3</sup>, Vincent. C. S. Lee<sup>4</sup>

<sup>1</sup> Department of Statistics, Chinese University of Hong Kong, Hong Kong, China.

<sup>2</sup> School of Management, Shandong University, Jinan, Shandong 250100, China.

<sup>3</sup> School of Mathematics, Shandong University, Jinan, Shandong 250100, China.

<sup>4</sup> Faculty of Information and Technology, Monash University, Melbourne 3800, Australia.

\* Corresponding author. Email: rongxiaoxia@sdu.edu.cn, vincent.cs.lee@monash.edu Manuscript submitted August 22, 2019; accepted February 12, 2020. doi: 10.17706/ijeeee.2020.10.4.312-320

**Abstract:** Virtual community (VC) arises rapidly and influences many aspects of human life styles in real world. Differentiated from traditional way to advertise products/services, VC also enables consumers to participate in interaction activities related to products via threads, learn greater insight about products in deep level while improve consumer loyalty. Most of the extant research did not emphasize or lack of effective methods on how to gain deep learning of product and explain the uniformity of users' importance in VC. In this paper, based on knowledge in complex network, generalised variance of degree in directed network is proposed to ascertain uniformity of directed network, which is an innovative methodology. Research conclusions can guide enterprises more in-depth understanding of the complex network theory and its application to social network analysis (SNA) with big data streams.

Key words: Virtual community, complex network, generalised variance of degree.

## 1. Introduction

Nowadays, data from multimedia social networks provide many information such as where users went and what they ate. Multimedia social data not only changed the culture of users' past experience, it also boosts our economy activity and shorten the social gaps among people. Virtual community (VC) is a special media as a result of "Internet+" [1], [2], arising rapidly to present functions about the products and as enabler for consumer to engage in product innovation [3]. Differing from traditional advertisement on TV or on-site manual post and the proliferation of advance posts in VCs enable consumers to carryout online query and acquire answers easily. At the commercial context, it is a good way for enterprise to save the cost of operating physical real shops. With the aid of VCs, enterprises are able to explore and exploit the opinion from consumers freely, and make relevant product improvements to add value of customer experience in the future [4]. From praxeology, consumers' first shopping experience may become their standards of reference for future purchase [5]. So, it is essential and desirable for enterprise to mine the key information from the VC's advance posts. At the technical view, VC contains special graph structure in which users (consumers) are regarded as nodes and their connections are edges (links), which indicate complex network is a useful tool to explore the properties.

Social, biology and technical network contain individuals with their interaction activities, which lead complex networks as an extension of the graph theory resulting to a practical method to analyse the

internal structure and dynamic process [6]-[8]. In Social Network Analysis (SNA), the interactions among the individuals have become an interesting part in information effective spreading and uncovering important nodes, communities and motifs [9]. In fact, nodes and edges represented different features of the data [10]. Networks were built according to intensive relations of nodes, and some properties such as "average degree" [11], "density of network" [12] and "eigenvector centrality" [13] were calculated to describe the key information of networks. The results can also help industries to govern the online service and optimize the costs.

Considering the practical applications to manufacturers, a question related with VC of enterprise are focused: How to judge whether the network holds the important macro- and micro- structures?

On this basis, this paper aims to explore the data from the methodology and case study from big data in Pollen Club. In methodology part, generalised variance of degree of directed network that proposed to judge the uniformity of the node importance in directed network. Thus, the consumer interaction network is established for case study. "Leaders" in network and special community are specified by PageRank and entropy method respectively.

The remainder of this paper is structured as follows. Section 1 introduces the research background and questions. Section 2 is related work on previous research methods. Section 3 describes the datasets of Pollen Club. Section 4 proposes a generalised variance of degree. Section 5 finds the "leaders" and core communities. Section 6 concludes and elaborates inspirations to enterprise management which are gained from case study.

#### 2. Related Work

Many studies related with VC were conducted by worldwide researchers. last 20 years has seen the rapid growth of the global computer network known as the internet [14]. Schrott and Beimborn [15] reported that VC changed people's life, which is widely accepted as important part of corporate knowledge management. VCs do release people'slocations even if the community and relevant topics are limited to a place [16]. Overall, there are mainly four kinds of methods concerning interactions in VC: Statistics, structural equation modelling, case studies, questionnaires and complex network.

In method of statistics, Khan *et al.* [17] analysed 1,922 brand posts from five different brands in three countries and ordinary least square and hierarchical moderation regression were used to test the important factors affecting consumers decisions on purchasing. Wan *et al.* [2] introduced LS-SVM innovatively into the study on VC's function in supply chains. Some researchers use structural equation modelling to explore the important interaction factors. Islam and Rahman [18] analysed the data using structure equation modelling through a questionnaire survey of 430 Facebook users. Simon *et al.* [19] developed and tested a conceptual research model empirically through structural equation modelling which is based on social impact theory, social identity theory and social exchange theory. Although the researches can specify factors that influence the VC, they ignore the topology information about community, which may root in the mechanism.

Other researchers conclude the factors from case studies and questionnaires. Grantham and Habel [14] used 9-point digital Likert scale questionnaire to investigate the degree to which participants perceived their VCs to exhibit each of the 10 target characteristics. Kilgour *et al.* [20] employed depth interviews initially, followed by questionnaires, and then content analysis was performed on 723 online media articles relating to social media marketing to identify semantic and conceptual relationships. The data of case studies are easily interfered by environment and samples of interviewers [21].

What is more, complex network is contained to gain insight of the VC. Park *et al.* [22] made use of the HITS algorithm to quantify the influence exerted by the opinion leaders on Twitter during the 2011 Seoul mayoral elections. Chen *et al.* [23] presented a simulated annealing algorithm to find important users by

optimising the user interest's concentration ratio in user groups. Chiang and Wang [24] conducted research on the interactive features of product-review networks with considering the out-degree centralisation, density, and microstructure of networks. However, these methods and traditional properties can not make a pre-judgement for the beingness of the important users ("leaders"). So, this paper proposes an innovative property in Section 4.

## 3. Data Sets

The dataset includes the replies of the corresponding main posts [25]. In detail, 129,362 users' information are retained containing 57,569 main posts and 886,087 replies.

Due to different roles that user played in the Pollen Club [26], users can be divided into three groups: Ordinary User Group (OUG), Intermediary User Group (IUG) and Enterprise User Group (EUG). OUG is the assemble of consumer who bought the product and register on Pollen Club to ask questions about product. IUG is responsible for helping OUG to solve problem under guidance of the EUG. And EUG contains employee, technical instructors, salesmen and propogandists of enterprise, which suggests different group undertake certain function in the Pollen Club. However, they also interact with each other, so data are firstly analyzed in the whole to specify key users in Pollen Club and consider the level information of users in community part. The relevant methods are showed in Section 4.

## 4. Methodology

Section 4 proposed one initiative property related with complex network to measure uniformity of directed network.

#### 4.1. Problem Statement

Variance can be used to demonstrate the distribution of the random variables in statistics. Previous studies were done to design the properties like "Variance of degree" [27] and "Generalised Variance of degree of network" [26] in undirected network and judge whether some "leaders" are in the network, who hold the special position among the nodes. Some examples of undirected network and directed networks:



Fig. 1. Examples of undirected network and directed network.

Compared with Fig. 1(a), the edges should have directions. In this paper, user b in Fig. 1(b) reply user c 3 posts which could not illustrate complete in Fig. 1(c). So generalised variance of degree of undirected network could not uncover this information. However there exists no research to definite the variance of the directed network. This property is defined in Section 4.2.

## 4.2. Generalised Variance of Degree of Directed Network

Generalised variance of degree of directed network is to specify uniformity of the node importance. Since the degrees of nodes in weighted directed network have two kinds containing generalised indegree and generalised outdegree. The generalised indegree of node i is the sum of the weight of arrows whose end point is i between their neighbors, denoted  $d_i^{gin}$ . Conversely, the generalised outdegree of node i is the sum of the weight of arrows whose start point is i between their neighbors, denoted by  $d_i^{gin}$ .

314

As a result, this paper defines generalised variance of degree in directed network in the following way:

$$G^{D}Var(G) = \frac{1}{n} \sum_{i=1}^{n} \left( d_{i}^{gin} - d_{i}^{gout} \right)^{2}$$
(1)

What is more,  $G^{D}Var(G)$  in Eq. (1) can degenerate into variance of degree of unweighted directed network showed in Eq. (2):

$$G^{D}Var(G) = \frac{1}{n} \sum_{i=1}^{n} \left( d_{i}^{in} - d_{i}^{out} \right)^{2}$$
(2)

Because  $d_i^{in} = d_i^{gin}$  under unweighted condition. Overall Eq. (2) is the general defemination of variance of degree in directed network.

Also,  $G^{D}Var(G)$  includes the variance of generalised outdegree of node *i* and generalised indegree of node *i*,  $G^{D}Var(G)$  can be decomposed in the followings (Eq. (3)):

$$G^{D}Var(G) = \frac{1}{n} \sum_{i=1}^{n} \left( d^{gin} - d^{got} \right)^{2}$$
  
=  $Var(d^{gin}) + Var(d^{gout}) - 2Cov(d^{gin}, d^{got})$  (3)

From the result showed in decomposition,  $G^{D}Var(G)$  excludes the covariance of generalised indegree and outdegree. It is reasonable that the great  $Cov(d^{gin}, d^{got})$  means the great mutually equal between each two nodes, which can make  $G^{D}Var(G)$  judge the difference among the nodes precisely.

To judge whether the network has "leaders", null model  $G^{D}(V', E', W')$  is introduced to make the rule. Generalised variance of degree of  $G^{D}(V', E', W')$  is  $G^{D}Var(G)^{zero}$  and the standard deviation of  $G^{D}(V', E', W')$  is  $\sigma[G^{D}Var(G)^{zero}]$ . Since generalised variance of degree of  $G^{D}(V', E', W')$  matches the Z distribution, which is proved to approximate normal distribution. In this paper, "leader" network is defined as: if the generalised variance of degree of G is bigger than the " $3 - \sigma$ " margin of null model  $G^{D}(V', E', W')$ ; that is  $GVar(G) > Mar^{D}(G) = E[G^{D}Var(G)^{zero}] + 3\sigma[G^{D}Var(G)^{zero}]$ , or "autonomy" network otherwise. From this definition, "leader" network has significant nodes named "leaders" controlling other nodes and influencing the generalised variance of degree, however the importance of "autonomy" network is relatively uniform.

#### 5. Results

Firstly, user interaction network G(V, E, W) is established in the following way: the nodes represent 129,362 users. If user *i* replied the post of user *j*, there would be arrow from node *i* to node *j*, of which the weight notes the frequency of reply. In symbol G(V, E, W), *V* is the set of all the users and *E* is the set of links (edges) among the nodes, with *W* noting the set of weights on links. A graph abut G is showed by Gephi in Fig. 2:

After calculating by Matlab according to the Eq. (3),  $G^D Var(G) = 382850.6910 \gg Mar^D(G) = 12.9252$ , which shows the network is a special network containing significant powerful nodes. So, PageRank is used to mine the "leaders" in G.

The thought of PageRank was established to rank the impotence of web page, which is determined by numbers and quality of pages which send a link to that page [28]. It is the same with interaction of people who has close relation with someone always replies to that person frequently. Since there are 129,362

315

nodes, so the network includes 12,936 "leaders". Due to the length of paper, table only list information about first 10  $PR_i$  in Table 1:

Table 1. Top 20 Nodes with High Lagenank value								
No	ID	$PR_i$	Group	No	ID	$PR_i$	Group	
1	92727	0.038713	IUG	6	59364	0.018902	OUG	
2	104364	0.03682	IUG	7	3807	0.016866	EUG	
3	5168	0.027376	OUG	8	5919	0.015837	OUG	
4	120317	0.022443	IUG	9	6050	0.015052	IUG	
5	119739	0.021215	EUG	10	6162	0.013146	EUG	

Table 1. Top 20 Nodes with High PageRank Value



Fig. 2. The users' interaction network.

It can be seen from Table 1 that OUG and IUG take a great proportion of users. So, in this online club, EUG is not the major part to participate in interaction. However, IUG play an important role in the leading members in OUG to discover the solution and beauty of the phone. Overall, the Pollen Club is very active, fascinating the OUG to explore and exchange opinions with others.

Moreover, sometimes the small communities assemble users with same interests. So, it is convenient to learn about the topics concerned, when it comes to a large scale of data. In another way, it can also inspire the enterprise to broadcast the product starting with which small group from analysis.

Due to difference in interaction among the users, network is composed of a series of weekly connected graphs. Through calculating the number of the nodes within the network, except for 2,746 isolated nodes, other comments two or more node. So the biggest component (noted  $G_1$ ) has over 120 thousand nodes, which is typical sample of G. Without loss of generality, this section separates the  $G_1$  into different communities, combing the PageRank value and level feature of the users to rank the communities. The Fig. 3(a) shows the separating results of communities, in which the nodes with same color stand for the same components:



Fig. 3. The separating results of the  $G_1$ .

It can be seen from the color that  $G_1$  contains many communities. Which community holds the special position physically? To solve this problem, "community network" is established by using a "big node" to represent the community and showed in Fig. 3(b). And community *I* has an arrow to community *J*, of which

the weight is the sum of all arrows' weights from community *I* to *J*, when the nodes within the community *I* send a link to that of community *J*. Because the topology positions of communities determine their function about the ability in spreading information in the network. So betweenness centrality property is used to calculate the number of the shortest path passing each community by which it can describe whether the communities hold the "bridge" positions.

However, there exists an important procedure to adjust the weight. In "community network", weight means the closeness between each two communities. In computing betweenness centrality property, weights stand for the cost of each two communities. So, the bigger weights mean the more closeness and the less costs. Weights are changed by Eq. (4), according previous analysis:

$$\overline{w_{IJ}} = 2 - \frac{w_{IJ} - min(w)}{max(w) - min(w)}$$
<sup>(4)</sup>

where min(w) and max(w) represent the minimum and maximum weight of each two communities in "community network". The range of  $\overline{w_{IJ}}$  is [1,2]. Apart from the position of network, an important community has a strong relation with the number of nodes which has high PageRank value. Following this, the betweenness centrality of communities, the average value of the PageRank value of nodes in separate community, and the number of nodes whose level belongs to OUG, IUG and OUG are calculated by Matlab. Finally, entropy method is introduced to established a judgement function to rank the communities.

Entropy method is an objective method to determine the weights of features, avoiding influence from subjective matters like AHP (Analytic Hierarchy Processing). Take *X* as data of the matrix of communities, features  $x_1 - x_3$  represent the number of the users in each community whose levels are in OUG, IUG and EUG respectively.  $x_4$  stands for the average value of PageRank values of nodes in communities and  $x_5$  is the betweenness centrality of communities. Since 5 features hold different data range, firstly *X* is normalized to  $X^{new}$  and entropy weight is calculated as  $\alpha_i$  in Table 2:

Table 2. The Weight of Features							
Features	x <sub>1</sub>	x <sub>2</sub>	X3	X4	x <sub>5</sub>		
Weight	0.3697	0.2388	0.0964	0.2113	0.0838		

What is more, an evaluation function J(x) is as following (Eq. (5)):

$$J(\vec{x_i}) = \sum_{j=1}^{5} \alpha_j \, x_{ij}^{new} \tag{5}$$

Since the thought of entropy method calculates the weights by divergence of features. So, in Table 2 there is a significant difference among communities' composition especially in number of users in OUG and IUG. At the same time, average PageRank values of nodes play an important role in discriminate the communities. Also, relevant small weight of the betweenness centrality indicate most of communities stay in the peripheral of the network. Finally, top 5 communities are listed in Table 3:

Table 3. Top 10 Important Communities								
No	ID	x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>4</sub>	x <sub>5</sub>	J(x)	
1	169	32480	59	120	$1.06 \times 10^{-5}$	3657.34	0.65	
2	0	35728	65	9	$7.09 \times 10^{-6}$	380.85	0.51	
3	3	36533	28	3	$6.93 \times 10^{-6}$	123.68	0.44	
4	1	6577	138	1	$5.63 \times 10^{-6}$	63.39	0.32	
5	9	16	0	0	$4.96 \times 10^{-5}$	0.00	0.21	

Table 3 figures out number 169 community holds the most important position. Although the number of

users in number 169 community approximates to that of the number 0 community, however with higher average value of PageRank value and betweenness centrality. What is more, number 169 community is a center of information, which is deserved for enterprise to analyze for broadcasting their information in the future. The values of evaluation function of other communities approximate to each other.

# 6. Conclusions

The consumer virtual community is explored by methodology and case study based on data from Pollen online club. In methodology part, generalised variance of degree of directed network is proposed to judge the uniformity of the node importance in directed network. Later, the consumer interaction network is established. Generalised variance of degree of directed network explore the uniformity of network. "Leaders" in network and special community are specified by PageRank and entropy method respectively.

Our new property related with complex network theory can be easily generalised into other fields. Generalised variance of degree of directed network can test the "leaders" ahead of time, when it comes to big data. Also results in this paper showed that consumer virtual community provides sufficient information where consumers and enterprise communicate with each other freely. Product manufacturers and services providers can provide better service and improve the product and service qualities according to the feedbacks from the web. Overall, enterprise should stimulate the IUG to keep on service for OUG to ease management pressure.

## **Conflict of Interest**

The authors declare no conflict of interest.

# **Author Contributions**

In this work, Zhen Zhang analyzed the data and wrote the major part of the paper. Qingchun Meng and other authors modified the logic and the language of this work. All authors had approved the final version.

## Acknowledgment

Zhen Zhang, Qingchun Meng are the co-first authors. This work was supported in part by the National Nature Science Foundation of China (NSFC) under Grant 71572096

## References

- Farseev, A., Lepikhin, K., Schwartz, H., Ang, E. K., & Powar, K. SoMin. ai: Social multimedia influencer discovery marketplace. *Proceedings of the 26th ACM International Conference on Multimedia (MM'18)* (pp. 1234-1236). New York: ACM Press.
- [2] Wan, X. L., Zhang, Z., Rong, X. X., & Meng, Q. C. (2016). Exploring an interactive value-adding data-driven model of consumer electronics supply chain based on least squares support vector machine. *Scientific Programming*, 8, 13.
- [3] Joo, J., Steinert-Threlkeld, Z. C., & Luo, J. Social and political event analysis based on rich media. Proceedings of the 26th ACM International Conference on Multimedia (MM'18) (pp. 2093-2095). New York: ACM Press.
- [4] Xiong, Y., Cheng, Z. C., Liang, E., & Wu, Y. B. (May. 2018). Accumulation mechanism of opinion leaders' social interaction ties in virtual communities: Empirical evidence from China. *Computers in Human Behavior*, 82, 81-93.
- [5] Ariely, D. (2008). Predictably Irrational. New York: HarperCollins.
- [6] Boccaletti, S., Latorab, V., Chavez, Y. M. & Hwang, D.-U. Complex networks: Structure and dynamics.

Physics Reports, 424(4-5), 175-308.

- [7] Newman, M. E. J. The structure and function of complex networks. SIAM. Review, 45(2), 167-256.
- [8] Newman, M. (2010). Networks: An Introduction. Oxford University Press, Oxford, UK.
- [9] Claudio, C., Santo, F., & Vittorio, L. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2), 591-646.
- [10] Roberts, S. G. B., Dunbar, R. I. M. T., Pollet, V., & Kuppens, T. Exploring variation in active network size: Constraints and ego characteristics. *Social Networks*, *31(2)*, 138-146.
- [11] Riyanto, Y. E., & Jonathan, Y. X. W. Directed trust and trustworthiness in a social network: An experimental investigation. *Journal of Economic Behavior & Organization*, *151*, 234-253.
- [12] Stark, T. H. The density of social networks moderates effects of intergroup contact. *International Journal of Intercultural Relations*, *55*, 133-147.
- [13] Houston, J. F., Lee, J., & Suntheim, F. Social networks in the global banking sector. *Journal of Accounting and Economics*, *65(2-3)*, 237-269.
- [14] Grantham, J., & Habel, C. An investigation of factors driving virtual communities. Proceedings of the Thirteenth Australasian User Interface Conference (AUIC '12) (pp. 91-92). Australian Computer Society Inc,. Australia: Darlinghurst.
- [15] Schrott, G., & Beimborn, D. Managing virtual knowledge networks: Topology and performance. Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work (pp. 198-204). New York: ACM Press.
- [16] Zhang, N. Simplifying cooperation with end users. *Proceedings of the 2015 on MobiSys PhD Forum* (pp. 21-22). New York: ACM Press.
- [17] Khan, I., Dongping, H., & Wahab, A. Does culture matter in effectiveness of social media marketing strategy? An investigation of brand fan pages. *Aslib Journal of Information Management*, *68(6)*, 694-715.
- [18] Ul, J., & Rahman, Z. The impact of online brand community characteristics on customer engagement. *Telematics and Informatics*, *34(40)*, 96-109.
- [19] Simon, C., Brexendorf, T., & Fassnacht, M. The impact of external social and internal personal forces on consumers' brand community engagement on Facebook. *Journal of Product & Brand Management*, 25(5), 409-423.
- [20] Kilgour, M., *et al.* The social media transformation process: Curating content into strategy. *Corporate Communications: An International Journal, 20(3),* 326-343.
- [21] Bernard, H. R., Killworth, P., Kronenfeld, D., & Sailer, L. The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology*, *13(1)*, 495-517.
- [22] Park, B., Lee, K., & Kang, N. The impact of influential leaders in the formation and development of social networks. *Proceedings of the 6th International Conference on Communities and Technologies*, (pp. 8-15). New York: ACM Press.
- [23] Chen, H., Cheng, X. Q., & Liu, Y. Finding core members in virtual communities. *Proceedings of the 17th International Conference on World Wide Web* (pp. 1233-1234). New York: ACM Press.
- [24] Chiang, C. H., & Wang, J. C. The impact of interaction networks on lurkers' behavior in online community. 1644-1656. IEEE.
- [25] Pollen Club. 2018. Pollen Club Forum: P10/P10 Plus Area. Retrieved from the website: https://cn.club.vmall.com/forum-2827-1.html
- [26] Meng, Q. C., Zhang, Z., Wan, X. L., & Rong, X. X. Properties exploring and information mining in consumer community network: A case of Huawei Pollen Club. *Complexity*, 2018, 19.
- [27] Liu, C. J. (2014). Community detection and analytical application in complex networks. Ph.D. Dissertation. Dept. Mathematics, Shandong Univ., Jinan. China: Shandong.

[28] Xing, W. P., & Ghorbani, A. A. Weighted PageRank algorithm. *Proceedings of Conference on Communication Networks & Services Research* (pp. 19-21). IEEE Computer Society Washington, DC.

Copyright © 2020 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (<u>CC BY 4.0</u>).



**Zhen Zhang** was born in China, 1993. He received his B.S. degree in 2016 and master degree in 2019. He is now with the Department of Statistics, Chinese University of Hong Kong, Hong Kong, China. His research interest is complex network, data analysis, and supply chain.



**Qingchun Meng** was born in China. He is now a professor with the School of Management, Shandong University, Jinan, China. And he is director of the Value Co-Creation Network Research Center of Shandong University, and associate dean of School of Management, Shandong University. Currently, he is mainly engaged in teaching and research in logistics and supply chain management, operations research and economic analysis.



**Xiaoxia Rong** was born in China. Since 1998, she has worked in the School of Mathematics of Shandong University. Her major is for operations research. Her research interests are mathematical planning, random planning and application. She was the main participant in the national operation research series teaching team and the excellent course operation research, and has won the second prize of national teaching achievement.



**Vincent. C. S. Lee** was born in Singapore. He is currently associate professor at Clayton School of IT, Monash University, Australia. He received Australia Federal Government scholarship to pursue PhD from 1988 through to 1991 at The University of New Castle, NSW, in Australia. In 1973 to 1974, he was awarded a joint research scholarship by Ministry of Defence (Singapore) and Ministry of Defence (UK) for postgraduate study at Royal Air Force College, UK in aircraft electrical and instrument systems. He was a visiting

academic to Tsinghua University in Beijing at the School of Economics and Management from Nov. 2006 to March 2007; and was also a visiting professor to Information Communication Institute of Singapore from July 1994 through June 1995.