An Empirical Investigation of Preferential Attachment Among Influential Members of a Large Artificial Community

JADERICK P. PABICO
jppabico@uplb.edu.ph
Institute of Computer Science, University of the Philippines Los Baños, Laguna
PHILIPPINES

Abstract – *One among the many questions in social network analysis is how links form among members of Internet-mediated social network (ISN), where most members are usually anonymous, while link formation (i.e., interactions) between members are facilitated only by non-personal communication technologies. Researchers offer preferential attachment (PA) as a possible mechanism that can explain the behaviour of link formation, not only for real-world communities, but for artificial communities, such as ISNs, as well. PA suggests that members choose to be linked with members characterized with many links who are considered "central" to the community. This is because it is believed that central members can be relied to as a channel, if not the source themselves, of information, of wealth, or of any other kind of currency that the community is using. In this paper, the link formation process of members of one large ISN was examined to look for empirical evidences of PA among members who were clustered together according to the order of magnitude of their number of links at the global level. Members whose initial number of links that totals only up to ten thousand exhibit the opposite of PA, while members whose initial number of links that sums greater than ten thousand exhibit PA. This means that the lower bound for initial links for PA, at least for this particular ISN, is 10,000. Additionally, for those members whose link formation follow the PA mechanism, the order of magnitude of the rate of increase in their number of links is proportional to the order of magnitude of their initial number of links.*

Keywords – preferential attachment, link formation, social network, artificial community, Internet-mediated society

I. INTRODUCTION

Since the start of the 21st century, the ubiquitous yet pervasive nature of the Internet has become an important aspect in the lives of the connected humans through the various communication services in the World Wide Web. These web services together with the humans who use them are collectively called Internet-mediated social network (ISN). Examples of ISNs that have gained global popularity, among many others, are Facebook® (www.facebook.com), Twitter® (www.twitter.com), LinkedIn® (www.linkedin.com), Google+® (plus.google.com), and Instagram® (www.instagram.com). These ISNs allow their members to publish personal information such as name, age, gender, civil status, place of residence, various personal media such as photographs and videos, and even the members' list of friends. Through this seemingly global publication platform, a person ***A*** that would otherwise not be aware of the existence of another person ***B***, who are for example geographically situated far away from ***A***, is now given the opportunity to meet ***B***, albeit in a "less" personal (or face-to-face) way. The capability to make known to the world that a particular person exists is probably one of the reasons why the ISNs have seen unprecedented growth in terms of usage and membership in recent years. The growth of membership of ISNs among various cultures in the world has become of interest to present-day scientists. Because of this, many growth models have been proposed with varying purposes from understanding the nature of ISNs, to how one can exploit the big data mined from them in marketing and in advertising.

The most common ISN growth model, among the many that exist, is the preferential attachment (PA) model proposed by Barabasi and Albert in 1999 [1]. In this model, when a person ***A*** joins an ISN, she* forms links to existing members $\boldsymbol{B}_1$, $\boldsymbol{B}_2$, ..., and $\boldsymbol{B}_n$ with probability proportional to the respective current number of links of $\boldsymbol{B}_1$, $\boldsymbol{B}_2$, ..., and $\boldsymbol{B}_n$. For example, if ***A*** wishes to be linked to $n = 4$ members of an ISN with $\boldsymbol{B}_1$ currently having $L_1 = 3$ links, $\boldsymbol{B}_2$ with $L_2 = 1$ link, $\boldsymbol{B}_3$ with $L_3 = 2$ links, and $\boldsymbol{B}_4$ with $L_4 = 0$ links, then ***A*** would have chosen to link with $\boldsymbol{B}_1$ with a probability equal to $P_1 = L_1 \times (\Sigma_{j\rightarrow 1..4}\, L_j)^{-1} = 1/2$, with $\boldsymbol{B}_2$ with probability $P_2 = 1/6$, with $\boldsymbol{B}_3$ with probability $P_3 = 1/3$, and with $\boldsymbol{B}_4$ with probability $P_4 = 0$. In general, the ith member $\boldsymbol{B}_i$ will be chosen by the newly arrived member ***A*** with a probability $P_i = L_i \times (\Sigma_{j\rightarrow 1..n}\, L_j)^{-1}$, where the symbol $\Sigma_{j\rightarrow 1..n}$ means summed over all indexes j from 1 through n, with n being the size of the community before the arrival of ***A***.

In ISNs whose links between members are deemed to be undirected, i.e., a member ***A*** having a link with a member ***B*** implies ***B*** having a link with ***A***, PA as presented above is already sufficient to describe the growth of the ISN. Many such real-world ISNs exist such as Facebook which requires ***B*** to "accept" a link request (or *friend request*) by ***A***. Accepting a link request makes ***A*** to be linked to ***B***, and at the same time makes ***B*** to be linked to ***A*** (or in the parlance of Facebook, ***A*** *becomes friend with* ***B*** and ***B*** *becomes friend with* ***A***). On the other hand, in ISNs whose links between members are deem to be directed, i.e., a member ***A*** having a link with a member ***B*** does not necessarily imply ***B*** having a link with ***A***, two important questions are needed to be asked first before one can

* The use of the female gender in this text is a choice of writing style, and depending on the context, both or either gender is meant without prejudice to the opposite one.

P-ISSN 2350-7756 | E-ISSN 2350-8442 | www.apjmr.com

understand link formation: (1) If ***A*** chooses to be linked, which current member reciprocate links, and to whom they reciprocate links with? And (2), given that reciprocal links already exist between two members, which link comes first? Under PA, the links are usually initiated by members whose current number of links are far way too small than most of the others. For example, many Twitter users "follow" a celebrity who are deemed influential in the community, but that same celebrity will not necessarily "follow" her "followers" who are practically unknown in the community. In the language of Twitter, to "follow ***B***" means to be linked to ***B*** in an unidirectional way.

Even though PA and similar other models have been proposed, few studies have attempted to validate these models from empirical observations of real-world, large, and highly dynamic ISNs. Moreover, for those who tried to validate the models with these ISNs, they encountered difficulty in using historical link addition records, where the time *T* of link formation between two members ***A*** and ***B*** is recorded. An intuitive reason for this is that during the time that these researchers were observing the ISNs' dynamics, the addition and deletion of links were highly "volatile." That is, before they can finish counting the number of links of a member ***A***, the number of links of ***A*** has already grown or shrunk, specifically (and more specially) for those ISN members whose number of links are extremely high. For example, in Twitter, the member with the highest number of follower links has about 40 million, and this is not yet counting the member's number of followed links. Thus, intuitively, even though one can count these links with a fast enough computer or cluster of collaborating computers, the exact number of links that one ISN member has at a specified time *T* can not be approximated confidently much so be counted exactly. In recent years, however, some of these ISNs have introduced the notion of "timeline" in their services, which provides their members a historical view of the members' interactions with other members of the ISN. In fact, Facebook and Twitter are the first two ISNs which provided their members with this kind of temporal view of their members' use of the service. Twitter has even provided an application-programmer interface (API) that allows computer programmers to mine the vast data that their members generate, perhaps to help them improve the system and to allow their third party business partners to exploit the community for marketing and advertising purposes. Advertising is one of the business models why these ISNs continue to thrive financially and improve their services [2].

In this study, an empirical dataset was collected to investigate whether the PA phenomenon exists in a large, globally popular, highly dynamic ISN Twitter. If PA does exist, then an answer may be provided to the question *how does the current number of links influence the speed by which one earn links from other ISN members*? By using the API provided at the developer's section of the ISN's website, the time-dependent growth of the number of links of some members was observed over a period of seven days, sampled every 15 minutes. It was found out from the analysis of the initial number of links that the sampled members can be grouped into cliques where each member's similarity is measured by how much number of links each one has at the start. Mathematically, members were considered in a clique if they share the same logarithms of the number of links. It was found out that members whose number of links is greater than 10,000 exhibit PA while those whose number of links is less than 10,000 exhibit the opposite. This means that the lower bound for PA in Twitter is 10,000 links. Moreover, it was also found out that for those members whose link formation follow the PA mechanism, the order of magnitude of the rate of increase in their number of links is proportional to the order of magnitude of their initial number of links.

II. Review of Literature

A. Origin of PA and other similar models

The PA phenomenon is an ubiquitous *rich-get-richer* mechanism for modeling the growth of complex communities usually found in nature [3]–[8]. The mechanism was first proposed (and named) by Barabasi and Albert [1] as a solution to the shortcomings of the random graph model (RGM) used by other researchers to explain the various phenomena observed in real-world communities. The RGM does not support the power-law distribution of the number of links of the community members. PA is the most recent of many names given to the mechanism of distributing a quantity (e.g., wealth, fame, credit, information, etc.) among members of a community according to how much these members already possess. It has been given such names in the past as *Yule* process [4], [6], cumulative advantage [9], rich-get-richer, *Matthew* effect [10]–[11], *Gibrat*'s rule on proportionate growth [3], [12], and *Zipf*'s law [13]–[14]. Most of these names were coined from what the researchers knew during the time of their research. The sociologist Merton [10]–[11], for example, coined the term Matthew effect from a verse in the Gospel of Matthew in the New Testament of the Holy Bible, which pertains to Jesus' "Parable of the Talents" that say

> *For whoever has will be given more, and they will have an abundance. Whoever does not have, even what they have will be taken from them.*
>
> – Matthew 25:29 (New International Version)

In this research effort, it is hypothesized that the Matthew effect is observable also in ISNs, particularly in the area of the number of "friends" one member has. In particular, it is hypothesized that the magnitude of how fast a member gains new friends is positively correlated to the magnitude of the initial number of friends that she has. For the purposes of the ISN under study, PA is formally defined as follows:

> **Definition 1**. *Preferential Attachment is a mechanism of link creation wherein the rate of increase in one's number of followers is proportional to how much one initially has.*

B. Mathematical formalization of PA

Mathematically, PA declares that the probability of a new member ***A*** linking (or initiating relationships) with an existing

P-ISSN 2350-7756 | E-ISSN 2350-8442 | www.apjmr.com

member ***B*** of the community is highly dependent on the centrality of ***B***, where centrality [15]–[16] is a measure of one's relative importance (or influence) within a community. That is, new members are more likely to form links with members who are deemed to be central (or influential) than with less central ones. This mechanism for community growth induces a degree distribution of the number of links of the members of the community described by the power-law distribution (Equation 1). The probability **Pr[***i*, *L***]** of the *i*th member having *L* links is equal to some normalizing constant β and a power function γ of *L* with normally (but not exclusively) $2 \leq \gamma \leq 3$. The range of values for γ has been found by many researchers to be true in most real-world communities [17].

$$\mathbf{Pr[}i, L\mathbf{]} = \beta \times L^{-\gamma} \qquad (1)$$

Other disciplines have also extensively used the power-law distribution to explain the dynamics of their observed data. Usually, these dynamics are based on the interrelationships between entities of their studies. For example, in biological sciences the power-law distribution explains the relationships among species as members of an ecological community, while the same distribution explains the relationships among computer hardware treating them as members of an information network, which is in turn considered as the community of hardware devices. Because the power-law distribution has been observed in various other communities, not only human communities but even ecological and computer communities, several modified PA models that take different aspects of the membership growth of the community have been suggested [18].

As early as 1925, Yule [4] published a stochastic preferential growth model to describe the uneven distribution of species among plant genera, which was later generalized by Simon [19] in what is now commonly called the Yule-Simon distribution. When this model is adapted to describe the growth of a community, the Yule-Simon process is defined with respect to a clique *C*[*L*] (i.e., a subset of that community) whose members are identified by their identical number of links *L*. The probability that a member of the clique will be linked by others is proportional to the abundance of links in that clique. More formally,

$$\mathbf{Pr[}i, C[L]\mathbf{]} = L \times N_L \times (\Sigma_{i \to 1..n}\, i \times N_i)^{-1}, \qquad (2)$$

where N_i is the number of members in the community whose number of links is *i*. Notice that the Yule-Simon distribution [19] generates an asymptotic probability function $P_j \sim j^{-(1+(1/(1-\alpha)))}$, where α is the ratio of the arrival rate of new members against the formation rate of new links. Notice further that Equations 1 and 2 are mathematically closely related when $\alpha = 1/2$. In this research effort, however, the members were deemed belonging to a clique if they have the same magnitude of number of initial links. Two members *x* and *y* belong to a clique if their respective number of links L_x and L_y are $\lfloor \mathbf{log_{10}}(L_x) \rfloor = \lfloor \mathbf{log_{10}}(L_y) \rfloor$, where the function $\lfloor z \rfloor$ means the largest integer less than *z*.

C. PA in dynamic communities

PA in dynamic communities is measured by calculating the rate ϕ_L at which a clique *C*[*L*] form new links during a small time interval Δ*T* [20]. This method has been extensively used to estimate PA in various relationships such as the links between scientific papers in communities called scientific citation networks, the links between scientists in communities termed as co-author and author collaboration networks, the links between web pages in the Internet, and the links between sex workers and their customers in a community called sexual networks, to name a few [21]–[32]. Calculating $\phi_{L,\Delta T}$ at time *T* is summarized in a function as follows:

$$\phi_{L,\Delta T} = (\Sigma_{i \to 1..n}\, \Delta L_i) \times (N_{L,T})^{-1} \sim \beta \times L^{-\gamma}, \qquad (3)$$

where ΔL_i is the number of new links that linked to members in clique *C*[*L*] during Δ*T* and $N_{L,T}$ is the number of members with *L* links at the start of time period *T*. Equation 3 should prove true for short time intervals Δ*T* during which the community should be observed to be in static (or steady-state) mode (or that the community has not gain new or lost old members, neither has any member gain new or lost old links). In other words, $N_{T+\Delta T} \approx N_T$.

For communities whose growth can be approximated by a linear function, the functional form of the asymptotic probability function can be determined from the γ exponent in Equation 3. The independent works of Krapivsky, *et al*. [33] and of Dorogovtsev, *et al*. [34] provide detailed examples of the approximation method for the linear form. According to the PA hypothesis, ϕ_L should increase monotonically with *L* (i.e., when $\gamma = 1$), and the linearity of PA is needed to generate a fat-tailed probability function with $P_K \propto L^{-\gamma}$. For exponents that are below linear (i.e., when $0 \leq \gamma < 1$), the probability function is a stretched exponential $P_K \propto L^{-\gamma} \times \exp(-(b_\gamma/(1-\gamma))L^{1-\gamma})$, where b_γ is a γ-dependent constant. There is a special situation where the preference among members is absent (i.e., $\gamma = 0$). In this case the rate ϕ_L becomes constant, and $P_L \propto \exp(-L)$ closely resembles the Poisson distribution in RGM. Finally, for $\gamma > 1$, the growth of the community leads to a behavior in which few very influential members are practically linked to all other members in the community.

D. The Twitter ISN

Twitter is an ISN that provides a microblogging service which enables its members to send and read 140-character limited text messages they called "tweets." A registered Twitter member can read and send tweets, while unregistered ones can only read them. Twitter members can access the service through an interface from a website, SMS, or mobile device application. The ISN was founded in 2006 by Jack Dorsey, Evan Williams, Biz Stone and Noah Glass. The microblogging service gained popularity worldwide that it hit a reported membership of 500 million in 2012. These members, and maybe more, generate about 340 million tweets daily, which made them one of the ten most-visited websites in the Internet. Their service has been described as the SMS of the Internet [35]. Because of this, researchers are now looking at tweets to

P-ISSN 2350-7756 | E-ISSN 2350-8442 | www.apjmr.com

predict the mood of a large community [36]–[37], or to "nowcast" possible contagion of flu symptoms [38]. Nowcast is the term used to forecast the next 15 minutes to one hour of a natural event such as rainfall.

E. Tweets, link types, and link formation mechanism

A member ***A*** in Twitter ISN may be linked through any of the two types of links by other members: followed and following. Member ***A*** is followed by member ***B*** if some tweets from ***A*** is read by ***B***. When ***B*** follows the tweets of ***A***, the Twitter interface automatically sends ***A***'s tweets to ***B***'s account, allowing ***B*** to read it, and then reply to it, "favorite" it, or "retweet" it. Replying to a tweet simply informs the originator ***A*** that one of her followers ***B*** is reacting to the tweet's contents. The reply may trigger a private discussion between ***A*** and ***B***. Making a tweet a "favorite" simply informs the tweet source ***A*** that her follower ***B*** is liking the content of the tweet. The number of favorites a tweet accumulate may indicate how useful to ***A***'s followers the tweet contents were, which will allow other followers to retweet it. The number of favorites somewhat provides a *cumulative advantage* to that specific tweet. Retweeting a tweet simply relays the exact message of the tweet from ***A*** through ***B*** and down to the followers of ***B*** [39]–[40]. In effect, the retweeting of tweets mimic the diffusion of the information contained in the tweet. The diffusion of information, as observed by other researchers, has similar dynamics as that of contagion in disease epidemics [35], [41].

III. Materials and Methods

A. Identifying influential Twitter members

A third party web service called Social Bakers® (www.socialbakers.com) provides the world's top Twitter members based on their respective number of followers (*F*), number of followings (*f*), and number of tweets (*t*). The web service also lists the Philippines' most popular Twitter members based on the same metrics *F*, *f* and *t*. Table 1 shows the global and the Philippine's respective top five Twitter members with their corresponding *f*, *F* and *F/f* ratio. Table 1 does not show *t* because it takes much more resources to query *t* than to query *F* or *f* for all members. Because of this, although not documented, it can be inferred that Twitter designed its database with "member" and "tweet" as separate entities (or objects), rather than the "tweet" object being an attribute of the "member" object. Intuitively, *t* can be seen as a derived attribute that can only be computed by querying the joint "member" and "tweet" objects, which requires much more computational resources. On the other hand, the values *F* and *f* are both inherent attributes of the "member" object, computing which only requires querying the "member" object, and thus are easy to obtain.

The global top five most followed Twitter members are mostly entertainers in the music industry, with the exception of the top fourth who is a known political world leader. The Philippines' top five are all female celebrities in the show business industry, except for the top third who is a known male comedian and TV reality show host. It can be seen from the table that the global top five outnumbered the Philippine's top five by a ratio ranging from 8.4 to 13.5 in terms of *F*. Intuitively, this means that at least there are 8.4 global followers for every one Filipino follower.

Table 1

Top 5 verified global and Philippine Twitter members with the most number of followers (*F*).
The data was gathered 31 December 2013 at 5:15pm (GMT+0800).

Rank	Name and Twitter Account Code	Number of Followings (*f*)	Number of Followers (*F*)	Followers-Followings Ratio
	Top 5 verified global Twitter members with the most *F*			
1	Katy Perry (@katyperry)	129	49,025,191	380,040
2	Justin Bieber (@justinbieber)	122,603	48,166,116	393
3	Lady Gaga (@ladygaga)	135,304	41,029,235	303
4	US President Barack Obama (@BarackObama)	654,859	40,764,315	62
5	Taylor Swift (@taylorswift13)	119	37,949,515	318,903
	Top 5 verified Philippine Twitter members with the most *F*			
1	Anne Curtis-Smith (@annecurtissmith)	1,276	5,826,248	4,566
2	Angel Locsin (@143redangel)	443	4,069,917	9,187
3	Jose Marie Viceral (@vicegandako)	531	3,443,572	6,485
4	KC Concepcion (@kc_concepcion)	616	3,150,778	5,115
5	Bianca Gonzales (@iamsuperbianca)	711	2,804,483	3,944

P-ISSN 2350-7756 | E-ISSN 2350-8442 | www.apjmr.com

B. Collecting link data

Using the command line computer program `twurl` [42], and together with the new Twitter API version 1.1 [43], the f and F of exactly 79 Twitter members were automatically recorded starting at 5:15pm of 31 December 2013, and every 15 minutes thereafter until 5:00pm of 7 January 2014. All time data are Philippine Standard Time (or GMT+0800). Of these 79 members, 22 are Filipino Twitter members, two of which are national political leaders, three are popular sportsmen, while the rest are prominent names in the entertainment and show business. The 22 Filipino members are composed of 11 males and 11 females, who were arbitrarily chosen because Social Bakers [44] has already verified that their identities are correct. That is, the members' published names in the Twitter community are owned by the respective same persons in real life, verified via official press releases, official websites, or personal contact. The remaining 57 members are non-Filipinos, most are prominent American entertainment industry names which are dominated by female singers. Other globally known names include US President Barack Obama, Microsoft owner Bill Gates, the Dalai Lama, and the Pope. The respective identities of these 57 members have been also verified by Social Bakers [44].

The specific Twitter API used to gather the personal information of members was `GET /1.1/users/show.json`, which queries the Twitter database specified by `user_id` or `screen_name` parameter. The `twurl` was ran on a ten-node cluster of personal computers, each with Intel Core i7-3632QM quad-core processor running the Scientific Linux v6.4 64-bit operating system with a 8GB random access memory. A cluster of computers was utilized to speed up the query process in an attempt to at least minimize the pitfalls introduced by the expected "volatility" of the Twitter ISN. Each 15-minute run was scheduled by Linux' `crontab` scheduler. The output of the query was recorded in a file named `results` wherein each line corresponds to the record of a Twitter member at that particular query time. All records were tab-delimeted and include the observation time T, Twitter account K, f, and F. The member's f and F data were observed for a total of at most 57,792 observations during the 7-day observation time. Since the recording of data started on the 31st of December 2013 at 5:15pm, and succeeding recordings happen every 15-minute interval, T was converted into an integer to simplify archiving of records. Thus, 5:15pm of 31 December 2013 was designated $T = 0$, while 15 minutes after (i.e., 5:30pm of the same day) was designated $T = 1$. The same conversion was done for all observation times T until 5:00pm of 7 January 2014, when $T = 671$.

C. Conducting the link analysis

The Kth member's time-dependent data on her number of followings $f_{K,T}$ and number of followers $F_{K,T}$ at specific observation time T were analyzed graphically for growth patterns. The members were clustered according to the magnitude of their $F_{K,0}$ (i.e., at $T = 0$). The $F_{K,0}$'s range from 6,580 to 49,025,191, a range that is clearly divided according to the order of magnitude. Thus, the members were clustered into four groups $\{G_i \mid i = 1, ..., 4\}$ with the first group composed of members whose respective $F_{K,0}$'s number into thousands and tens of thousands (G_1), the second group with $F_{K,0}$'s that number into hundreds of thousands (G_2), the third group with $F_{K,0}$'s that number into millions (G_3), and the fourth group with $F_{K,0}$'s that number into tens of millions (G_4). G_1 has 4 members, G_2 has 9 members, G_3 has 42 members, and G_4 has 24 members. The $F_{K,T}$ gain or loss of each member at each succeeding time observation $T+\Delta T$ was computed as

$$\Delta F_{K,T} = F_{K,T} - F_{K,0}, \forall T > 0, \quad (4)$$

and then averaged across all members of the same group. That is, $\mathbf{mean}[\Delta F_T]_i = (\Sigma_{K \in Gi}\, \Delta F_{K,T}) \times |G_i|^{-1}$, where $|G_i|$ is the size of the ith group. To quantify the spread of ΔF_T at each group, the standard deviation $\mathbf{std}[\Delta F_T]_i$ was likewise computed. The data pair (T, $\mathbf{mean}[\Delta F_T]_i$) was fitted into the monomial

$$Y_i = \alpha_i \times T^{\varphi}, \quad (5)$$

where Y_i is an estimate for $\mathbf{mean}[\Delta F_T]_i$. The coefficients α_i and φ_i were estimated by converting Equation 5 into a one-degree binomial using logarithms (i.e., $\mathbf{log}\ Y_i = \mathbf{log}\ \alpha_i + \varphi_i\ \mathbf{log}\ T$), and conducting a simple linear regression on this conversion. Patterns of the amount of φ_i each group has were observed. This was done to test the following hypotheses:

> $\boldsymbol{H}_0$: (Null Hypothesis) *The PA mechanism is not evident in the dynamics of the ISN under study.*
>
> $\boldsymbol{H}_1$: (Alternate Hypothesis 1) *The PA mechanism is evident to all members across groups and that the φ_i has a proportional magnitude as the $F_{i,0}$.*
>
> $\boldsymbol{H}_2$: (Alternate Hypothesis 2) *The PA mechanism is evident only to members in groups G_2 through G_4 and that members in group G_1 either do not exhibit it or exhibit a preferential detachment* [45] *mechanism.*

The null hypothesis $\boldsymbol{H}_0$ is accepted if all alternate hypotheses $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ are rejected. However, if $\boldsymbol{H}_0$ is rejected, then any of the alternate hypotheses is accepted. $\boldsymbol{H}_1$ says that all sampled members of the ISN, regardless of their respective memberships to cliques, exhibit PA. Statistically, this means that the estimate of φ across all groups is greater than zero, while its value is significantly different from zero at 5% confidence level. $\boldsymbol{H}_2$ means that all members in groups G_2 through G4 exhibit PA while those who belong in G_1 do not. Statistically, the estimate $\varphi_i > 0$ and $\varphi_i \neq 0$ at 5% confidence level, $\forall\ i>1$, and that the estimate $\varphi_1 \leq 0$ at the same confidence level. Those whose $\varphi < 0$ are exhibiting a mechanism called *preferential detachment*, a term introduced by Miritello [45], and is the exact opposite of PA.

P-ISSN 2350-7756 | E-ISSN 2350-8442 | www.apjmr.com

IV. Results and Discussion

A. Twitter members with the most $F_{K,0}$

Figure 1 shows the respective $F_{K,0}$ of members deemed to be influential in the Twitter ISN. A member ***A*** is considered influential in her community if F_A is large compared to that of the other members of the community [46]. Figure 1a shows the respective $F_{K,0}$ of the 4-member G_1 composed of Caps Cop, Fart Robot, Ms. Kim Kardashian, and Robot J. McCarthy with their respective $F_{K,0}$ of 23,775, 10,411, 6,580 and 10,527, and **mean**$[\Delta F_T]_1$ = 12,823. Caps Cop is the known alias of the member who publicly campaigns against indiscriminate use of capital letters in Twitter posts, while Fart Robot is known worldwide for collecting one-liner quotes related to flatulence. Ms. Kardashian is a controversial American actress and Mr. McCarthy is a fiction writer known for espousing conspiracy theories.

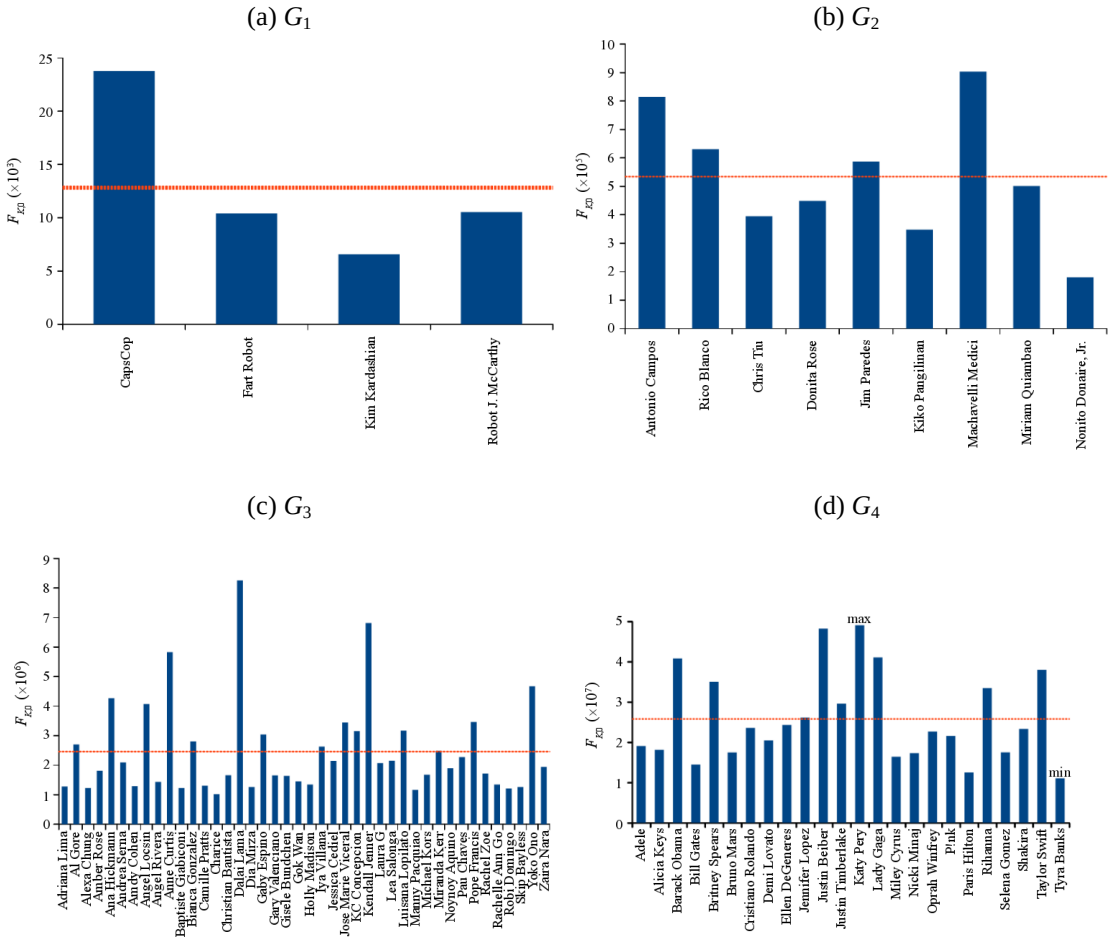

Fig. 1. The respective $F_{K,0}$ of deemed-influential Twitter members by cluster group. The blue bars represent the $F_{K,0}$s while the orange dashed lines represent the **mean**$[\Delta F_T]_i$ for G_i. Note that the respective vertical axes between groups differ by orders of magnitude.

Figure 1b presents the individual $F_{K,0}$s of G_2's nine members composed of Antonio Campos, Rico Blanco, Chris Tiu, Donita Rose, Jim Paredes, Senator Francisco Pangilinan, Machavelli Medici, Miriam Quiambao, and world-class champion boxer Nonito Donaire, Jr. Their $F_{K,0}$s average to **mean**$[\Delta F_T]_2$ = 534,370 with a maximum of 903,135 and a minimum of 180,593. Mr. Campos is a celebrity personality of spanish descent while Mr. Medici is a novelist and TV commercial model. Messrs. Blanco and Paredes are Filipino music artists, while Messrs. Tiu and Donaire are influential Filipino sports personalities. Ms. Rose is known in the Asian region as a video jockey for a popular video channel and Ms. Quiambao is an international beauty titlist. Sen. Pangilinan is a nationally-elected member of the upper chamber of the Philippine Congress.

Figure 1c shows the respective $F_{K,0}$s of the 42-member G_3. The group has a minimum $F_{K,0}$ of 1,015,737 belonging to Ms. Charice Pempengco, a maximum $F_{K,0}$ from the Dalai Lama, and a **mean**$[\Delta F_T]_3$ = 2,460,484. G_3 is composed of 25 female celebrities and 17 male personalities. Some notable personalities who are included in this group are former U.S. Vice President Al Gore, professional boxing's only 8-division world champion Manny Pacquiao, Philippine President

P-ISSN 2350-7756 | E-ISSN 2350-8442 | www.apjmr.com

Benigno "Noynoy" Aquino III, and John Lennon's spouse Yoko Ono.

The 24-member G_4 and its members' respective $F_{K,0}$'s are shown in Figure 1d with **mean**$[\Delta F_T]_4$ = 25,843,915, minimum of 11,018,866 from Ms. Tyra Banks, and maximum of 49,025,191 from Ms. Katy Perry. As of January 2014, Ms. Perry is world's Twitter user with the most number of followers. The group is composed of a world-influential political leader in U.S. President Barack Obama, an equally-influential technology and business leader Microsoft founder Bill Gates, an internationally-known soccer player Mr. Cristiano Ronaldo, 19 American female show business celebrities, and three American male show and music business personalities, namely Bruno Mars, Justin Bieber, and Justin Timberlake.

B. Link growth patterns of selected Twitter members

Figure 2 shows the respective $T \times F_{K,T}$ patterns $\forall T$ of some randomly selected K. Selected were U.S. President Barack Obama who is clustered to G_4, the Pope who belongs to G_3, Filipino world-class champion boxer Nonito Donaire, Jr. who represents G_2, and controversial American actress Kim Kardashian who is in G_1. Due to space constraints, these four were selected just for showing their individual patterns but the rest of the analysis will consist of the summarized values across all cluster members. In Figure 2, notice that the four patterns should not be visually compared to each other because their respective $F_{K,T}$ values differ by orders of magnitude over the other. For example, the $F_{K,T}$ of Pres. Obama is one order of magnitude greater than that of the Pope, and three orders of magnitude greater than Ms. Kardashian's. Figure 3, on the other hand, shows the respective $T \times \delta F_{K,T}$ patterns for the same selected K as above. The symbol $\delta F_{K,T}$ should mean the periodic gain (or loss if negative) by the Kth member. More formally,

$$\delta F_{K,T} = F_{K,T+1} - F_{K,T}, \forall T > 0, \quad (6)$$

where periodicity is the 15-minute observation interval. Notice how Equation 6 differ from Equation 4. As in Figure 2, the trends in Figure 3 must not also be compared to each other because their values differ by orders of magnitude.

President Obama's $F_{K,T}$ pattern shows an almost linearly increasing trend (Figure 2a) except for T≈200 and T≈300 where he lost approximately 11,000 to 12,000 followers (Figure 3a). He lost about 11,000 at T≈200, steadily gained about 10,000 until T≈300, lost practically the same number at T≈300, and then linearly gained followers after. Pres. Obama gained about 85,000 followers more at the end of the 7-day observation period. The times T≈200 and T≈300 were particularly important to Pres. Obama's $F_{K,T}$ trend because these are the times that coincide with the first day of implementation of the much-talked about Obamacare or the Affordable Care Act [47]–[48].

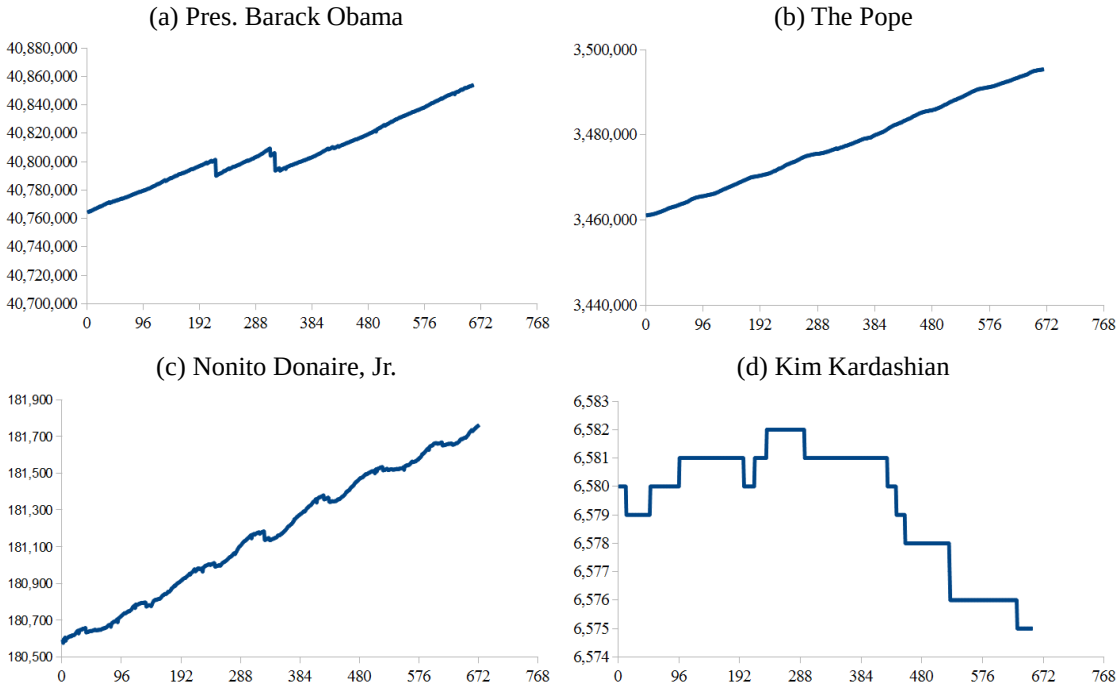

Fig. 2. The $T \times F_{K,T}$ patterns of some selected Twitter members representing their cluster group: (a) Barack Obama for G_4, (b) the Pope for G_3, (c) Nonito Donaire for G_2, and (d) Kim Kardashian for G_1. The interval T = 96 stands for one day.

P-ISSN 2350-7756 | E-ISSN 2350-8442 | www.apjmr.com

The Pope's $F_{K,T}$ pattern shows a monotonically linear increasing trend with T (Figure 2b). The Pope steadily gained about 35,000 more followers with loses up to 200 followers in one day at T≈300 (Figure 3b). This translates into a daily average of 5,000 linking rate for the Pope. Notice from Figure 3b that the link gain pattern $\delta F_{K,T}$ of the Pope seems to follow a "seasonal" trend, where one seasonal cycle roughly coincides with the middle of the day (i.e., "diurnal" trend) under the Philippine Standard Time (PST).

Filipino boxing champion Donaire's $F_{K,T}$ pattern shows a linearly increasing tendency coupled with a seemingly "diurnal" trend (Figure 2c). This diurnal trend can be visually confirmed by looking at his $\delta F_{K,T}$ trend pattern (Figure 3c), which resembles that of the Pope but at the next lower $\delta F_{K,T}$ magnitude. Mr. Donaire gained about 1,100 more followers at the end of the observation period.

Because the rate of growth of Ms. Kardashian's $F_{K,T}$ is slow compared to the three others, her pattern seems to show a step-wise rise and fall (Figure 2d) showing maxima of one follower gained and one follower lost once every several contiguous observation periods (Figure 3d). During the 7-day observation period, though, Ms. Kardashian gained a total of four followers but lost eight of them for a net loss of four followers. This roughly translates into an average of one follower being lost in two days.

C. Link growth rate and PA

Figure 4 shows an $F_{K,0} \times$ **mean**$[\Delta F_T]_i$ plot across all groups. The plot visually shows the mean change in the number of followers each ISN member has or the growth rate of their respective number of followers. The PA mechanism, according to Definition 1, suggests that the rate of increase in one's number of followers (ΔF_T) is proportional to how much one initially has ($F_{K,0}$). A plot that is drawn in the first quadrant suggests that such mechanism exists, i.e., the **mean**$[\Delta F_T]_i$ monotonically increases as $F_{K,0}$ is increased. In fact, the trend line drawn in Figure 4 is mathematically expressed as

$$\textbf{mean}[\Delta F_T]_I = 1.37\times10^{-6}\,(F_{K,0})^{0.92}. \qquad (7)$$

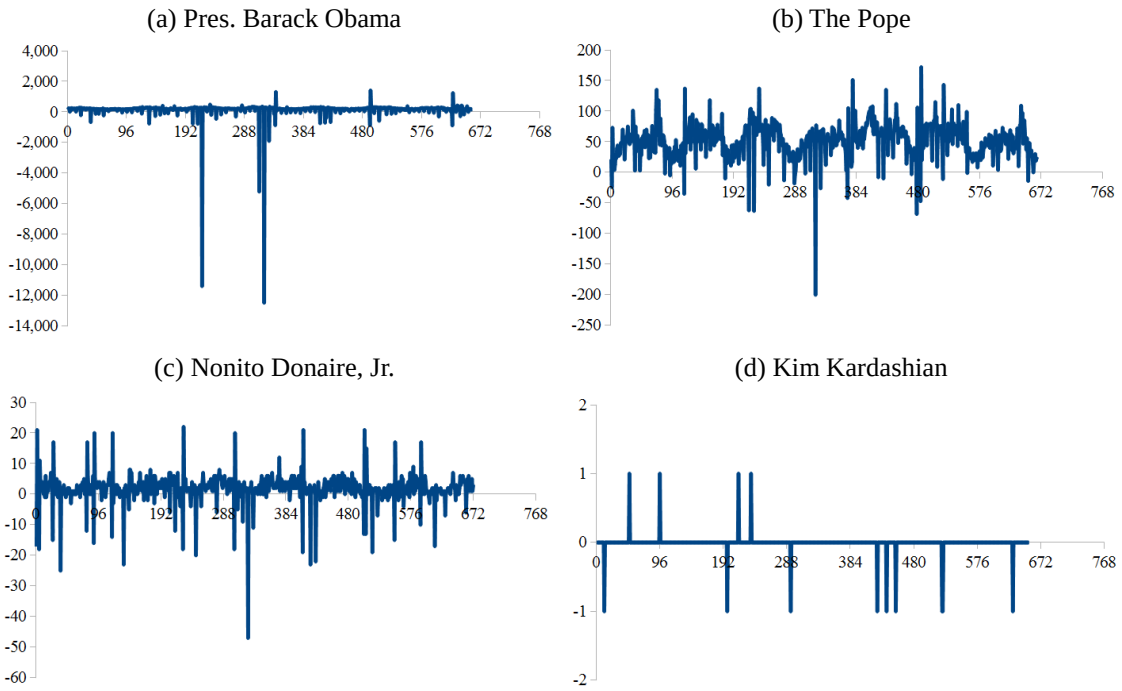

Fig. 3. The $T \times \delta F_{K,T}$ patterns of the same selected Twitter members as that of Fig. 2. (a) Barack Obama, (b) the Pope, (c) Nonito Donaire, and (d) Kim Kardashian. The interval T = 96 stands for one day.

The coefficient α and power φ constants were each statistically determined to be different from zero at 5% confidence level, while the trend has a coefficient of determination R^2 = 0.72. R^2 suggests that 72% of the data points can be explained by Equation 7. Definitely, the two extreme data points beloging to two members in G_3 with higher than normal **mean**$[\Delta F_T]_i$ are not explained by Equation 7. These two data points respectively belong to Mr. Angel Luis Rivera (c), a CEO of a clothing company in Puerto Rico with **mean**$[\Delta F_T]_i$ = 26.57 and to Mr. Baptiste Giabiconi (d), currently the world's highest paid male model with **mean**$[\Delta F_T]_i$ = 15.50. The two members with the highest $F_{K,0}$ are Ms. Katy Perry (a) and Mr. Justin Bieber (b) with **mean**$[\Delta F_T]_i$ = 21.57 and **mean**$[\Delta F_T]_i$ = 21.51, respectively. Even though Mr. Rivera has the highest **mean**$[\Delta F_T]_i$, this is not enough to overtake Ms. Perry as the member with the highest $F_{K,0}$ in the near future. Interestingly, if both their respective **mean**$[\Delta F_T]_i$ remain constant, Mr. Rivera will have the same $F_{K,0}$ as Ms. Perry in about 272 years, an impossibility with the current average lifespan of humans.

Since it was statistically found out that $\alpha \neq 0$ and that $\varphi \approx 1.00$, then regardless of the value of α, $\log(\textbf{mean}[\Delta F_T]_i) \approx \log(F_{K,0})$, suggesting that $F_{K,0}$ is positively

P-ISSN 2350-7756 | E-ISSN 2350-8442 | www.apjmr.com

related to **mean**$[\Delta F_T]_i$ via a proportional growth in the order of magnitude. Since the trend encompasses all groups G_1 through G_4, does this mean that hypothesis $\boldsymbol{H}_1$ is already accepted? Caution is advised to eagerly accept $\boldsymbol{H}_1$ at this point knowing that $|G_1| = 4$ only and $R^2 = 0.72$. It is possible that at least most members in G_1 have ΔF_T that are beyond what Equation 7 can explain (i.e., data points from G_1 could also be potentially outliers).

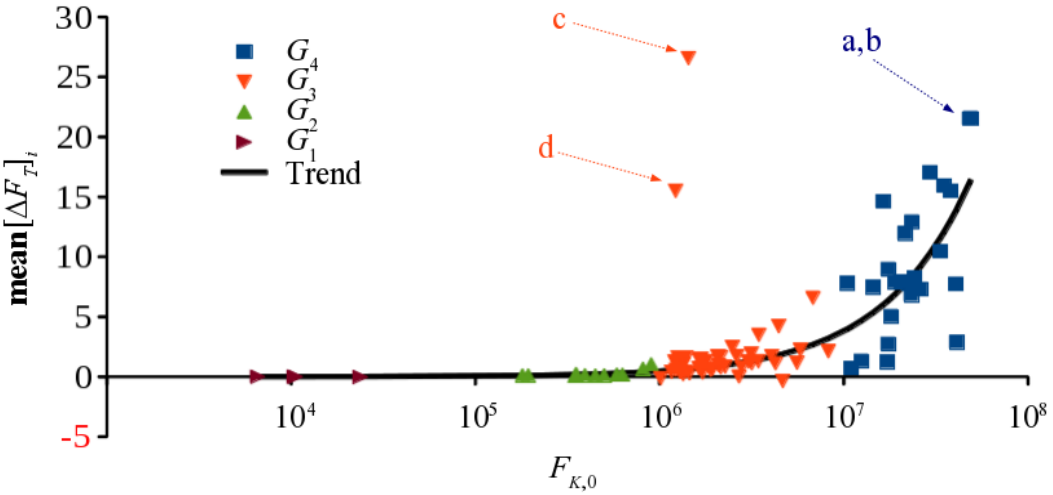

Fig. 4. The $F_{K,0}$ × **mean**$[\Delta F_T]_i$ plot across all groups.
Interesting data points discussed in the text are those of
Ms. Katy Perry (a), Mr. Justin Bieber (b), Mr. Angel Rivera (c), and Mr. Baptiste Giobiconi (d).

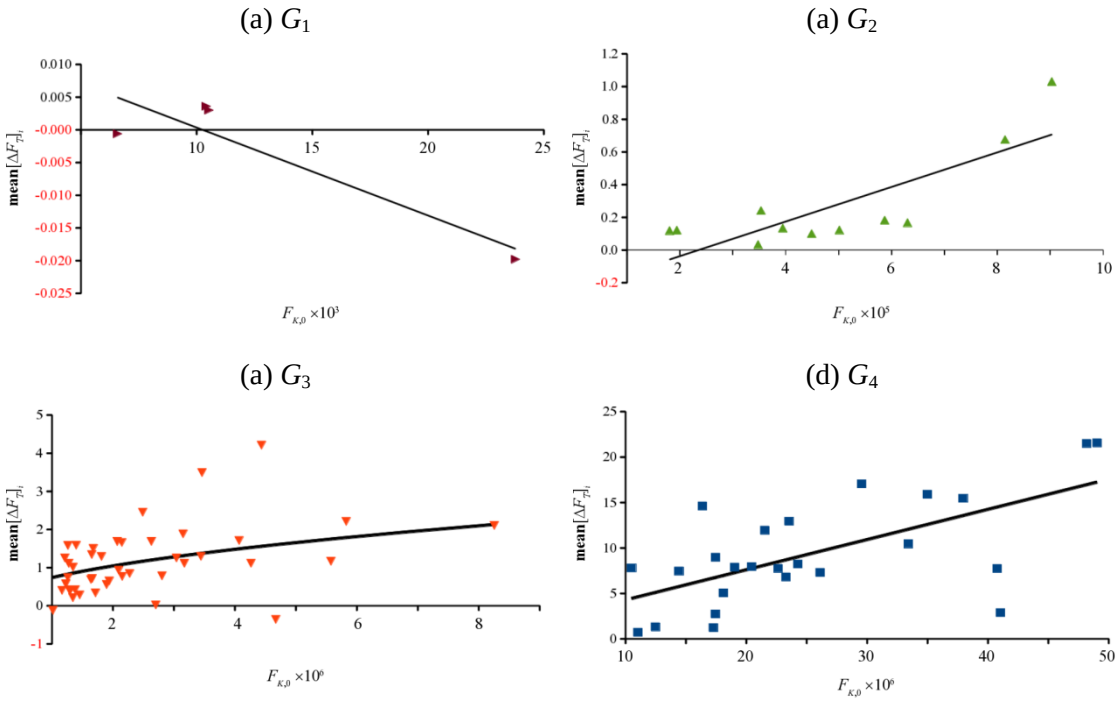

Fig. 5. The $F_{K,0}$ × **mean**$[\Delta F_T]_i$ plot per group.

Figure 5 shows the $F_{K,0}$ × **mean**$[\Delta F_T]_i$ plot per group. G_1, G_2, G_3, and G_4 have respective trend lines shown in Equations 8 through 11. All groups follow a linear trend except for G_3 which follow an exponential trend. The respective intercepts and slopes of the linear trend, as well as the coefficient and power of the exponential trend (Equation 10), were statistically found to be significantly different from zero at 5% confidence

P-ISSN 2350-7756 | E-ISSN 2350-8442 | www.apjmr.com

level. Note that the power coefficient transforms into a slope coefficient if the logarithm of both sides of Eqution 10 is to be taken. The slope and power coefficients of Equations 9 through 11 are positive which means that the rate of increase of the number of followers in groups G_2 through G_4 are likewise positive. This suggests PA for these groups. The rate of increase of the number of followers in group G_1 is negative, which suggests a mechanism that opposes PA. These findings allow one to accept the hypothesis $\boldsymbol{H}_2$ and reject both $\boldsymbol{H}_0$ and $\boldsymbol{H}_1$. The acceptance of $\boldsymbol{H}_2$ allows one to suggest that the lower bound for PA in the Twitter ISN is 10,000. Any member whose $F_{K,0}$ < 10,000 is expected to exhibit the opposite of PA.

$$\textbf{mean}[\Delta F_T]_1 = 0.01 - 1.35\times10^{-6}\,(F_{K,0}) \quad (8)$$

$$\textbf{mean}[\Delta F_T]_2 = -0.25 + 1.06\times10^{-6}\,(F_{K,0}) \quad (9)$$

$$\textbf{mean}[\Delta F_T]_3 = 7.04\times10^{-4}\,(F_{K,0})^{0.50} \quad (10)$$

$$\textbf{mean}[\Delta F_T]_4 = 0.95 + 1.37\times10^{-7}(F_{K,0}) \quad (11)$$

V. Conclusion and Extensions

This paper shows an empirical investigation of PA mechanism in the growth of the number of links of members in the Twitter ISN. Analysis shows that PA exists for members whose $F_{K,0}$ is at least 10,000. The rate of increase of $F_{K,T}$ at any time T is positive if $F_{K,0} \geq 10{,}000$, while the opposite is true for $F_{K,0}$ < 10,000. Thus, it is expected that PA is exhibited in Twitter ISN by those members whose $F_{K,0}$ is at least 10,000. Additionally, it was also found out that for those members whose link formation follow the PA mechanism, the order of magnitude of the rate of increase in their number of links is proportional to the order of magnitude of their initial number of links. Thus for members of Twitter that obey PA, it is expected that if two members' respective $F_{K,0}$'s differ by w order of magnitude, then their respective **mean**[ΔF_T]'s will also differ by w order of magnitude.

Two investigative works are already underway as extensions of this research endeavor. These works aim to answer the following questions:

1. *Does the Kth member belongs to cliques from her L_K*? In this work, not only the number of links (F or f) of K is counted but also each of these links is identified. By identifying K's followers or followings, one can infer whether K reciprocates her followers (or that K's followings reciprocate her). That is, if K is being followed by a member B, does she also follow B in return? If so, how many such reciprocated links exist in her F or f? For all members B_i with reciprocated links with K, do they also have reciprocated links with each other? The existence of such proves that cliques exist within K's links.
2. *How does the diffusion of tweets differ from members whose $F_{K,0}$'s and* **mean**[ΔF_T]*'s differ by orders of magnitude*? Forking from the work of [41], it is hypothesized that the rate of diffusion of tweets from members with high $F_{K,0}$'s and **mean**[ΔF_T]'s is faster than the those with low $F_{K,0}$'s and **mean**[ΔF_T]'s. Experiments are being conducted to statistically verify this hypothesis with empirical data sets.

Acknowledgements

This research effort is funded by the Institute of Computer Science (ICS) of the University of the Philippines Los Baños (UPLB) under the research programs *Structural Characterization and Temporal Dynamics of Various Natural, Social and Artificial Networks in the Philippines* and *Evaluating Approaches for Modeling and Simulating the Structure and Dynamics of Artificial Societies*.

The 10-node computer cluster that was used in this endeavor belongs to ICS' Research Collaboratory for High-Performance Computing.

References

[1] A.-L. Barabasi and R. Albert. (1999). *Emergence of scaling in random networks.* **Science** 286(5439): 509-512.

[2] BBC News. 2013. *Twitter plans stock market listing*. **BBC News** 12 September 2013 issue.

[3] R. Gibrat. 1931. **Les Inegalites economiques**. Ph.D. Thesis. Universite de Lyon.

[4] G.U. Yule. 1925. *A mathematical theory of evolution, based on the conclusion of Dr. J.C. Willis, F.R.S*. **Philosophical Transactions of the Royal Society B** 113(402-410):21-87.

[5] D.G. Champernowne. 1953. *A model of income distribution*. **The Economic Journal** 63(250):318-351.

[6] H.A. Simon. 1957. **Models of Man: Social and Rational**. John Wiley & Sons: New York, pp 301.

[7] D.J. de Solla Price. 1976. *A general theory of bibliometric and other cumulative advantage processes*. **Journal of the American Society for Information Science** 27(5):292-306.

[8] L. Hebert-Dufresne, A. Allard, V. Marceau, P.A. Noel, and L.J. Dube. 2011. *Structural preferential attachment: Network organization beyond the link*. **Physical Review Letters** 107(15):158702-158706.

[9] D. Dannefer. 2003. *Cumulative advantage/disadvantage and the life course: Cross-fertilizing age and social science theory*. **The Journals of Gerontology, Series B: Social Sciences** 58(6):S327-S337.

[10] R.K. Merton. 1968. *The Matthew effect in science*. **Science** 159(3810):56-63.

[11] R.K. Merton. 1988. *The Matthew effect in science II: Cumulative advantage and the symbolism of intellectual property*. **ISIS** 79(299):606-623.

[12] J. Sutton. 1997. *Gibrat's legacy*. **Journal of Economic Literature** 25:40-59.

[13] G.K. Zipf. 1949. **Human Behavior and the Principle of Least Effort**. Addison-Wesley.

[14] B. Jiang and T. Jia. 2011. *Zipf's law for all the natural cities in the United States: A geospatial perspective*. **International Journal of Geographical Information Science** 25(8):1269-1281.

[15] M.E.J. Newman. 2005. *A measure of betweeness centrality based on random walks*. **Social Networks** 27(1):39-54.

[16] S.P. Borgatti and M.G. Everett. 2006. *A graph-theoretic perspective on centrality*. **Social Networks** 28(4):466-484.

[17] B. Graham. 2013. **Nature's Patterns: The Art, Soul, and Science of Engaging Her**. Amazon Digital Services:New York, pp 120.

[18] S.N. Dorogovtsev and J.F.F. Mendes. 2003. **Evolution of Networks: From Biological Nets to the Internet and WWW**. Oxford University Press: New York, pp.

[19] H.A. Simon. 1955. *On a class of skew distribution functions*. **Biometrika** 42(3-4):425-440.

[20] H. Jeong, Z. Neda, and A.-L. Barabasi. 2003. *Measuring preferential attachment for evolving networks*. **Europhysics Letters** 61(4):567-572 (arXiv:cond-mat/0104131).

[21] M.E.J. Newman. 2001. *Scientific collaboration networks I: Network construction and fundamental results*. **Physical Review E** 64:016131-(1-8).

[22] M.E.J. Newman. 2001. *Scientific collaboration networks II: Shortest paths, weighted networks, and centrality*. **Physical Review E** 64:016132-(1-7).

[23] A.-L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. 2002. *Evolution of the social network of scientific collaborations*. **Physica A** 311: 590-614.

[24] M.E.J. Newman. 2006. *Mixing patterns in networks*. **Physical Review E** 67:026126.

[25] B.F. de Blasio, A. Svensson, and F. Liljeros. 2007. *Preferential attachment in sexual networks*. **Proceedings of the National Academy of Sciences of the United States of America** 104(26):10762-10767.

[26] J.P. Pabico. 2008. *Inferences in a virtual community: Demography, user preferences, and network topology*. **Philippine Information Technology Journal** 1(2):2-8.

[27] C. Bird, E.T. Barr, A. Nash, P.T. Devanbu, F. Filkov, and Z. Su. 2009. *Structure and dynamics of research collaboration in computer science*. In **Proceedings of the 9th SIAM International Conference on Data Mining**, pp. 826-837.

[28] J.P. Pabico. 2009. *Social network analysis of research collaboration among Filipino agricultural scientists*. In **Proceedings of the 9th ISAAS-Philippines National Convention and Annual Meeting** (in CDROM).

[29] J.P. Pabico. 2010. *The network structure of scientific research collaboration of agricultural engineers in the Philippines*. In **Proceedings of the Joint 8th International Agricultural Engineering Conference and Exhibition, 60th Philippine Society of Agricultural Engineers Annual National Convention, and 21st Philippine Agricultural Engineering Week** (in CDROM).

[30] J.P. Pabico. 2010. *Authorship patterns in computer science research in the Philippines*. **Philippine Computing Journal** 5(1):1-13/

[31] J.P. Pabico and J.R.L. Micor. 2009. *Structural analysis of the collaboration network of Filipino scientists*. In **Proceedings of the 2nd UPLB CAS Student-Faculty Research Conference**, NCAS, UPLB, Laguna.

[32] J.P. Pabico and J.R.L. Micor. 2013. *Ang social network sa Facebook ng mga taga-Batangas at ng mga taga-Laguna: Isang paghahambing*. **Asia Pacific Journal of Multidisciplinary Research** 1(1):138-150.

[33] P.L. Krapivsky, S. Redner, and F. Leyvraz. 2000. *Connectivity of growing random networks*. **Physical Review Letters** 85(21):4629-4632.

[34] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. 2000. *Structure of growing networks with preferential linking*. **Physical Review Letters** 85(21):4633-4636.

[35] L. D'Monte. 2009. *Swine flu's tweet tweet causes online flutter*. **Business Standard** 04 February 2009 issue.

[36] T. Lansdall-Welfare, V. Lampos, and N. Cristianini. 2012. *Nowcasting the mood of the nation*. **Significance** 9(4):26-28.

[37] V. Lampos, T. Lansdall-Welfare, R. Araya, and N. Cristianini. 2013. *Analysing mood patterns in the United Kingdom through Twitter content*. Technical Report (arXiv:1304.5507).

[38] V. Lampos and N. Cristianini. 2012. *Diagnosing flu symptoms with social media*. **Natural Hazards Observer** 36(4):7-9.

[39] R. Needleman. 2007. *Newbie's guide to Twitter*. **C|Net News** 15 March 2007 issue.

[40] K. Bodnar. 2011. *The ultimate glossary: 120 social media marketing terms explained*. **HubSpot** 30 December 2011 issue.

[41] A.C. Salvania and J.P. Pabico. 2010. *Information spread over an Internet-mediated social network: Phases, speed, width, and effects of promotion*. **Philippine Information Technology Journal** 3(2):15-25.

[42] M. Molina and E. Michaels-Ober. 2011. **Twurl**. (github.com/twurl).

[43] Twitter, Inc. 2014. **The Twitter Developer API v1.1**. (developer.twitter.com)

[44] Social Bakers. 2013. **Top Twitter users by country and by followers**. (www.socialbakers.com)

[45] G. Miritello, R. Lara, M. Cebrian and E. Moro. 2013. *Limited communication capacity unveils strategies for human interaction*. **Scientific Reports** 3:1950 (doi: 10.1038/srep01950).

[46] M. Kimura, K. Saito, R. Nakano, and H. Matoda. 2010. *Extracting influential nodes on a social network for information diffusion*. **Data Mining and Knowledge Discovery** 20(1):70-97.

[47] J. Easley. 2014. *President Obama's approval rating jumps 5 points as millions sign up for Obamacare*. **Politicus USA** 1 January 2014 issue.

[48] C. Weaver. 2014. *Health Law's uneasy launch: Affordable Care Act faces hurdles*. **The Wall Street Journal** 01 January 2014 issue.