Modeling Content Hotness Dynamics in Networks

Jagadish Ghimire, Mehdi Mani and Noel Crespi
Wireless Networks and Multimedia Service Department
Institut TELECOM SudParis, France
{*firstname.lastname*}@it-sudparis.eu

***Abstract*—Different kinds of networks at different levels of system design have evolved in the last decade, mainly riding on top of global Internet. Regardless of the type of the network, these networks can be viewed as content sharing and distribution network. Understanding the popularity dynamics of the contents, termed here as Content Hotness, is useful in many ways including the characterization of the workload as well as the system design and evaluation. Despite the fact that popularity skewness among contents has been well studied, the temporal dynamics of popularity of a given content has not been studied extensively. We attempt to propose a discrete time Markov chain (DTMC) model to model such content level hotness dynamics. We focus in two realistic scenario and see how such model can be used to represent the temporal variation in the content popularity.**

I. Introduction

Essentially, there are two aspects of content popularity (which is termed as content hotness in this paper). The first aspect deals with the characterization of the distribution of popularity among different contents in a network on a momentary basis. In other words, it deals with the fact that different contents achieve a different overall popularity. Many models including Zipf [8], stretched exponential [3] are suggested to represent such popularity-skewness among different contents. The other aspect of content popularity rather deals with the dynamics of the popularity of a given content over time. It is a well-observed phenomenon that even a content during its lifetime enjoys different level of popularity in time resulting in a phenomenon that we term here as "Content Hotness Dynamics". We see that this aspect of content popularity (the hotness dynamics)has not been studied extensively as opposed to the study of popularity distribution among contents. There however are few work including [7] [6] and [2] which focus in the content level hotness dynamics. Our work also deals with the hotness dynamics of the contents, also often referred to as the temporal dynamics of content popularity.

Before looking at what has already been done in this context, we first point out briefly the importance of understanding content hotness dynamics in networks. Knowledge of popularity skewness and its dynamics can be exploited for understanding and evaluating different information networks like the caching mechanism in Cache-systems, P2P networks and even the IPTV systems. Such modeling of popularity can be used in workload characterization and also in prediction [7]. We also see that proper modeling of popularity dynamics can help us in doing better caching decisions, better system dimensioning as well as better system design. Like many other important characteristics, hotness dynamics is an important characteristics of any content network.

Now, we explain what has been done in the previous work relating to temporal dynamics of hotness. [7] empirically studies the change in popularity of some selected contents in two different kinds of networks (Youtube and Digg). They find that the contents in Youtube appear to have a slower decay so that the views are distributed over time where as the contents in Digg (which essentially are news-like contents) decay faster in popularity. On the other hand, [6] attempts to model the popularity dynamics of an IPTV channel using a stochastic model called mean reversion model. This however is limited by the assumption that such stochastic model work only when any intermediate changes are supposed to settle to a long term equilibrium value like in financial system (e.g. stock). However, for a given content with a finite life-time, this assumption will not be valid in general. Thus, this model can not be used to model the hotness dynamics of contents. [2] attempts to come up with a closed-form expression for the hotness dynamics of content parameterized in such a way that the cumulative popularity can either be represented as an exponential decay function or a power-law function. It is however unknown how different network and content properties relate to the model parameters. Because of this fact, it is unclear how this model can be used to represent different network structures, content advertisement mechanisms and other factors that affect the hotness dynamics. We clearly see that in order to be able to model the hotness dynamics, one should also be able to understand and incorporate the network and content properties that affect the hotness dynamics. Considering this more general philosophy in mind, in this work-in-progress, we intend to put forward the model that we are working on.

To be more precise on the quantitative descriptions to follow in the following section, we define hotness of a content as follows. Hotness of any content i at any time t is a physical quantity proportional to the rate of request coming to that content at a particular point in time.

II. General Hotness Model based on DTMC

In a network of N nodes, a content c is originated at any of these nodes. Each node, aware of the availability of the content, and interested in it takes (downloads) the content. In this context, the content, with time, gets propagated in the network. The speed with which it propagates and other various related attributes depend heavily on the true mechanism of the content propagation in the network. There are many models

existent in the real and conceptual networks. In general, this involved complex notions including advertisement mechanism, content discoverabilty mechanism, topology of the underlying network and the content sharing principles.

Our model attempts to summarize the content's propagation in the network as a Discrete Time Markov Chain where a state x ($1 \leq x \geq N$)represents the number of nodes that have "viewed" the content. A transition probability P_{ij} from state i to state j represents the probability that the content, currently "viewed" by i nodes would be "viewed" by total j nodes after the DTMC time-step T_{obs}. In this work, we model the content propagation mechanism in the network approximately by a DTMC with the state space $X(= 1 \cdots N)$ together with the transition probabilities $P_{ij}(i, j \in X)$. Below, we will illustrate more on the estimation of the transition probabilities.

A. Definitions

The hotness of a content at a state x of the corresponding DTMC, represented as $h(x)$ is defined as a function of the current state. It is given as follows:

$$h(x) = \sum_{j=x+1}^{N} P_{i,j}(j - x) \tag{1}$$

This is proportional to the rate of requests coming for the content when it is in state x of its DTMC. This is in accordance to the conceptual meaning of hotness.

Now, we represent $H(n)$ as the hotness of the content in time nT_{obs}. It is governed by the state of the DTMC after the given time (represented as n transitions) and thus is a r.v.. Let $P_{i,j}^{(n)}$ represent the probability that the DTMC in state i will be in state j after n transitions. Then, r.v. $H(n)$ has a value $h(x)$ with a probability $P_{0,x}^{(n)}$ for all $0 \leq x \leq N$. $P_{0,x}^{(n)}$ can simply be obtained using Chapman-Kolmogorov formula if $P_{i,j}$ are known.

$$P\{H(n) = h(x)\} = P_{0,x}^{(n)} \tag{2}$$

Then, the expected hotness after a time nT_{obs}, represented as $\overline{H(n)}$ is given as follows.

$$\overline{H(n)} = \sum_{i=0}^{N} P_{0,i}^{(n)} h(i) \tag{3}$$

Now, an interesting parameter of content hotness, called the lifetime of the content is defined as the time after the origin of the content, when the content is distributed to every users in the network. Consider it as T_{span}. This is a r.v. with the following P.M.F.

$$P\{T_{span} = n\} = P_{0,N}^{(n)} \tag{4}$$

Then, the expected value of T_{span} represented as $\overline{T_{span}}$ is given as follows.

$$\overline{T_{span}} = \sum_{x=0}^{\infty} x P_{0,N}^{(x)} \tag{5}$$

This general DTMC model thus allows us to express the expected hotness of a content as a funtion of time. Using the expected hotness as a function of time ($\overline{H}(n)$), we can compare the hotness of two contents at a given point in time. Moreover, using the comparision of the expected span of the content, we can also compare how fast the content saturates in the network. For two contents with the same N, the one with smaller value of $\overline{T}_{span}$ can be considered a hotter content in average. Below, we will see two different scenario of the application of the model.

III. Scenario 1: Hungry Node Model (Information Spreading Mechanism)

In this section, we try to model a common phenomenon in social and other networks arising primarily because of the incomplete network structures.

1) Idea: In a social network, a user (interchangeably called as node) is connected to a subset of other users. Normally, these networks rely on the concept of sharing. For example, if I have a content, I share it among my friends (one hop neighbors in the equivalent graph). Because of this sharing, a subset of nodes become aware of the existence of the content whereas the rest of the network are unaware of the availability of the content. In this scenario, we introduce the concept of "Hungry Nodes". At any point in time, if the DTMC is in state x (i.e. x nodes have already viewed it), a total of $N(x)$ nodes are aware of the availability of the content including those who already "viewed" it. This means that there are $n(x) = N(x) - x$ number of nodes that are aware of the availability of the content and have not viewed the content yet. These nodes are considered to constitute of a Hungry Node Set. All the other nodes in the network are unaware of the content availability. The manner in which $n(x)$ changes in average depends largely upon the topological property of the network.

For a network of $N = 100$ nodes, we have calculated statistically the average value of $n(x)$ for a regular graph with average degree =3 and another network with same number of links but obtained to result in irregular structure (see our algorithm [4] for converting this regular graph to the irregular graph). The results show that between regular graphs and the shuffled graph (which follows a preferential attachment), in the regular graph the rises in $n(x)$ is slower. In both cases, however, $n(x)$ rises to a peak and starts to fall until $n(x) = 0$ for $x = N$. See figure 1. This is intuitively clear to see that topology affects the change in $n(x)$ with x.

A. p: The probability of content viewing

At any state x, we set that there will be $n(x)$ hungry nodes. Now another parameter to introduce in this context is the probability p that any node aware of the availability of the content will "view" the content in the next time-step of DTMC. Considering that this probability is independent and constant for all, we can now express the state transition probabilities in terms of p and $n(x)$ as described below.

For any state i, the probability of going to a state j in next transition for all $j < i$ is zero. Also, the probability of going

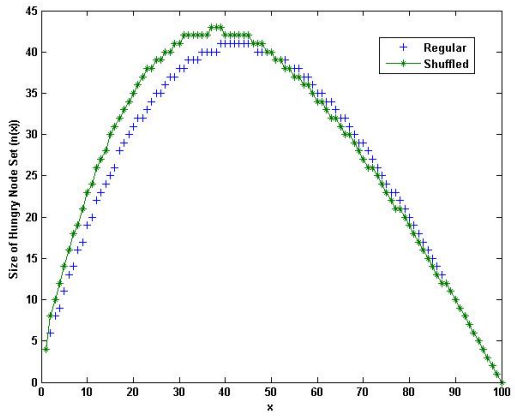

Fig. 1. Mean Size of Hungry Nodes versus the State of DTMC

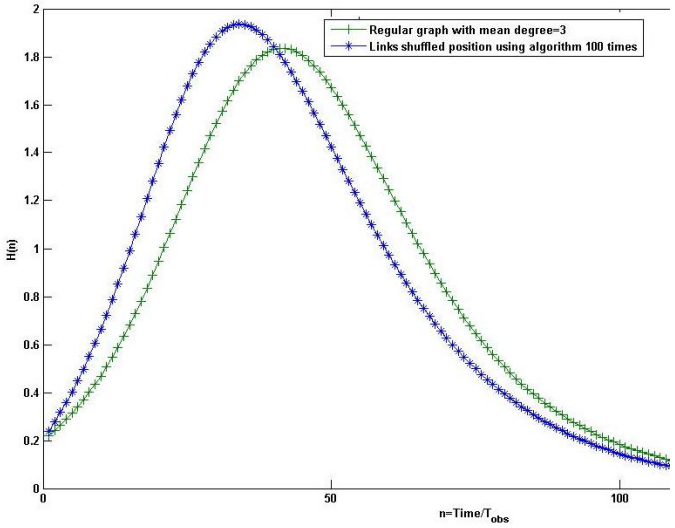

Fig. 2. Expected Hotness versus time

from state i to j in next transition for all $j > N(i)$ is also zero. The probability of this state transition from i to j for $i \leq j \leq N(i)$ is a binomial random variable parameterized at $n(i)$ and p. So we obtain the following transition probabilities for a basic hungry node model.

$$\begin{aligned} P_{ij} &= 0 \text{ For } j < i \\ P_{ij} &= 0 \text{ For } j > N(i) \\ P_{ij} &= \binom{n(i)}{j-i} p^{j-i} (1-p)^{n(i)-j+i} \text{ For } i \leq j \leq N(i) \quad (6) \\ N(i) &= i + n(i) \end{aligned}$$

Thus together with the knowledge of $n(i)$ and p, we can use the expressions for P_{ij} into the definitions of hotness from equations 2 and 3 to obtain the expected hotness profile of a network. We have used the $n(i)$ values for the regular graph and the shuffled version of the graph (shown in figure 1) and setting $p = 0.05$, we obtain the expected hotness versus time for these two kinds of graph. See figure 2 The hotness profile shows that under such model, hotness first rises to a peak and then falls back until finally it disappears. This nonlinear profile can be somehow mapped into three regions of content life-cycle: *Rise in popularity*, *peak popularity* and *fall in popularity*.

This model can be interesting to see how network topology affects the expected hotness of the content. Moreover, a beforehand knowledge of p and $n(i)$ can be used to predict the hotness of a content after a given time. The prediction of hotness has many interesting applications.

IV. Scenario 2: Hotness model based on Influence

In the Hungry Node Model, we assumed that p, the probability that a node will "view" the content in a given time-step if it is aware of the availability of the content, is constant. Keeping this constant however means that there is no phenomenon of Influence. However, an influence mechanism exists in many real networks because of different factors. Since p represents the promptness of a node to "view" the content, its value is thus affected by the influence mechanism going on in the network. Normally, influence can ba a very subjective notion and hard to model. But there are attempts ([1], [5]) to model the influence mechanism. No matter what the actual underlying influence mechanism is, it affects the value of p. In fact, the more people "view" the content, those who have not viewed it tend to be affected positively (though it can be a negative effect also) so that their p value tends to increase. This overall tendency or promptness to view a file, thus can be considered as the function of current spreading of the content in the network, i.e. x. Thus, in this section, we define p as state dependent $p(x)$. As an example, if there are more people watching the content, in the due course of time, the content becomes more valuable for the new ones who do not watch it. Similar phenomenon is seen among Youtube videos also. In literature, different models are proposed to model such influence mechanism. We do not attempt to go into specifics of such influence mechanism.

Also, in order to decouple the effect of information-spreading (scenario 1) from the effect of influence mechanism in the hotness dynamics of the content,we assume that all nodes are aware from the beginning , the availability of the content. In other words, $N(i) = N$ for all i. This will help us see solely the effect of influence on the hotness dynamics.

In this case, the state transition probabilities are represented as function of $p(i)$ as shown below. This makes the model an easy to use homogeneous DTMC.

$$\begin{aligned} P_{i,j} =& \binom{N-i}{j-i} p(i)^{j-i} (1-p(i))^{N-j-i} \text{ for } i \leq j \\ =& 0 \text{ for } i > j \end{aligned} \quad (7)$$

Using our model, we find out the expected hotness for

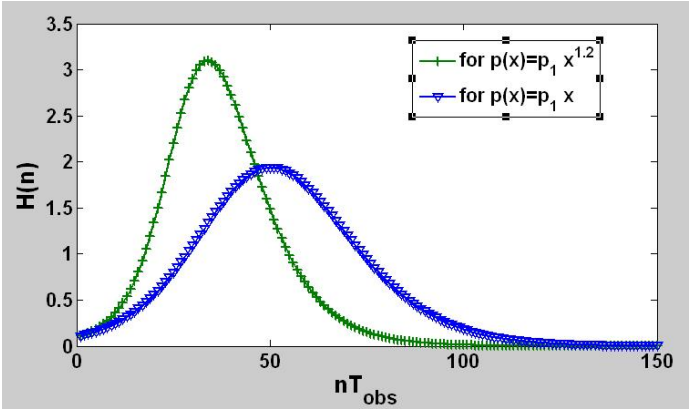

Fig. 3. Hotness for two different influence mechanisms, $p_1 = \frac{1}{1000}$

two cases: (a) $p(x)$ is linearly increasing with x (b) $p(x)$ is increasing as $x^{1.2}$. We can see in figure 3, how influence mechanism affects the hotness profile: both the peak hotness as well as the duration of the content. These preliminary results are interesting. Even a slight difference in the underlying influence mechanism results in a huge difference in the hotness dynamics of the content. This model can be used to capture such effects of influence mechanisms on the spread of contents and their temporal popularity variations. To the best of our knowledge, this work has not been dealt with in the past literature.

The specifics on how $p(x)$ changes with states depend upon the underlying Influence Model and the Topological Structure and also the advertisement network. Here in the entire work, we set that there is no explicit advertisement mechanism. All the content availability information is disseminated via the network propagation. We believe that hotness modeling (and thus prediction) should consider these structures and properties of the network and the users.

A. Constant p case: what are the examples?

We consider a case in which the probability that the content will be accessed by a user in the observation interval is constant, regardless of the current state of the DTMC. i.e. $p(x) = p$. This is a very basic model and is very specific. The scenario of such a model would be the system where the "interest of users on the content" does not increase with the distribution of the content. This could typically be a system, like:

- Systems with no influence or no "word-of-mouth": Here, the content is accessed (by a search in a P2P network or a request to a content server in a C/S network) independently to the spread of the content in the network by other nodes. This does not concern the raise in content discoverabilty as well as the social dimension resulting in influences.

B. What about the finite life span of the content

In [7] and motivated by this work, in the other work, authors have focused in establishing the fact that some contents

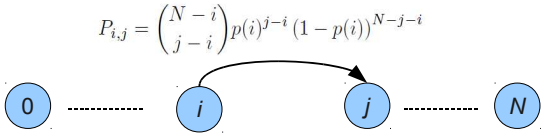

Fig. 4. A useful model for Content Hotness

(semantically categorized as news contents) have a finite life span because they loose relevance after a certain time. For such contents, the previous experimental studies shows that their popularity growth and decay are abrupt. Our model can also be used to model such finite life-span content. In that case, $p(x)$ is not only state dependent, it is also time dependent. Then $p(x)$ is represented as $p(x,t)$. This means that the resulting DTMC is no more homogeneous. Such a non-homogeneous DTMC, in our view, represents the general model for hotness dynamics of contents in a network. The simplest limited life-span model now introduces another parameter called T_{life} and $p(x,t)$ is represented as $p(x) \times [U(t) - U(t - T_{life})]$ where $U(t)$ is an unit step function.

V. Discussions and Future Direction

This work is just a starting point for modeling the content level hotness dynamics. We believe that being a simple and yet a general model, this can be used and improved to perform interesting studies relevant to content hotness. We can take real network topologies and well known influence mechanisms to map the experimental data of different content networks and then tune the parameters to predict the popularity of the content in such networks. We can also study the effect of different network design options in the hotness profile and then tune the system design practices. We believe that a model as general as this can be improved and adapted for different scenario.

References

[1] C. Asavathiratham, S. Roy, B. Lesieutre, and G. Verghese. The influence model. *Control Systems Magazine, IEEE*, 21(6):52 –64, dec 2001.

[2] Zlatka Avramova, Sabine Wittevrongel, Herwig Bruneel, and Danny De Vleeschauwer. Analysis and modeling of video popularity evolution in various online video content systems: Power-law versus exponential decay. *Evolving Internet, International Conference on*, 0:95–100, 2009.

[3] Lei Guo, Enhua Tan, Songqing Chen, Zhen Xiao, and Xiaodong Zhang. The stretched exponential distribution of internet media access patterns. In *PODC '08: Proceedings of the twenty-seventh ACM symposium on Principles of distributed computing*, pages 283–294, New York, NY, USA, 2008. ACM.

[4] http://sites.google.com/site/shufflealgorithm/.

[5] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, New York, NY, USA, 2003. ACM.

[6] Tongqing Qiu, Zihui Ge, Seungjoon Lee, Jia Wang, Qi Zhao, and Jun Xu. Modeling channel popularity dynamics in a large iptv system. In *SIGMETRICS '09: Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems*, pages 275–286, New York, NY, USA, 2009. ACM.

[7] Gábor Szabó and Bernardo A. Huberman. Predicting the popularity of online content. *CoRR*, abs/0811.0405, 2008.

[8] G. K. Zipf. Selective studies and the principle of relative frequency in language, 1932.