Measuring Internet Eco-system Using User-centric Data

Joon Han
Dept. of Sociology, Hallym University,
joonhan@sun.hallym.ac.kr

Chan-ung Park
Dept. of Sociology, Dongguk University,
chanung@dongguk.edu

In the present paper we view the Internet as an eco-system in which Internet sites compete with each other for visitors' attention. Based on an ecological perspective, we analyze the relationships among Internet sites based on the socio-demographic attributes of Internet users. The Internet users select Internet sites based on their interests and preferences, which are shaped by their socio-demographic attributes such as age, education, class, and gender. The socio-demographic spectrum of site visitors constructs a socially embedded web of Internet sites competing and cooperating for more visitors. To the extent that people develop their preferences and interests through socialization processes, those Internet users with similar characteristics are likely to visit similar sites, thus developing homophily between users and sites. Starting from an assumption of social embeddedness of the Internet, we have two goals in this paper: first, we will present two measures of niche width and the extent of niche overlap of Internet sites; and, second, we will analyze how the extent of niche overlap affects the performance of Internet sites through competition and mutualism among them.

In order to test ecological ideas against empirical data on the Internet, we use information collected through user-centric measurement of Internet since our method of measuring niches of Internet sites requires socio-demographic data of Internet site visitors. The panel data are provided by "Internet Metrix," a for-profit research organization specialized in Internet related research.

We measured the Internet site niche based on the Internet users' demographic attributes. We started by securing information on the distribution of visitors' demographic variables such as age, gender, years of schooling, and occupation for each site in our sample. Internet sites are then compared pairwise in terms of similarity of visitors' demography. At the same time, we calculated visitor overlap for each pair of Internet sites. With a total of n sites, we get the measure of similarity in visitors' attribute and the measure of visitor overlap for $n\times(n-1)$ pairs. In this manner, we measure the relative distance among Internet sites in terms of their visitors' similarity or overlap. We relied on following two indices to apply the concept of niche to the relations between Internet sites.

First, the niche overlap index was developed by Miller McPherson (1983) to measure the niche width of voluntary associations. We followed McPherson's suggestion and extracted descriptive statistics of visitors' age and years of schooling for each site. We then calculated each site's niche width as the mean $\pm$ (1.5×standard deviation) for each dimension and measured the common area through pairwise comparison of two sites' niche width. Second, the index of visitor overlap was calculated. This index measures the relative nearness of two sites more directly by the degree of two sites sharing the same visitors. This method of measurement is similar to deriving quasi-network from the joint involvement data in network analysis. After calculating the number of visitors shared by two sites, we normalized the index by dividing with the squared root of two sites' number of visitors. The index of visitor overlap ranges from 0 to 1: from no overlap between two sites to complete overlap.

Based on the two indices of overlap between Internet sites, niche overlap and visitor overlap, we drew niche maps of Internet ecosystem using multidimensional scaling. On the maps, the broader the sites' niche width is, the closer it will be to the center of the map. At the same time, the sites located close to each other would share larger number of visitors or attract visitors with similar demographic background. The graphs coming from the multidimensional scaling mark sites using the relative distances among them, and, therefore, the axes, Y or X, do not have meaning by themselves.

We applied regression models to our data to estimate the competition or mutualism effects of niche overlap and visitor overlap on the performance of Internet sites. In the regression models, we included average values of niche overlap and visitor overlap against other sites. We used independent variables measured in June to predict

the dependent variable measured in October. To distinguish between competition effect and mutualism effect of overlap, we separately measured overlap among the sites within the same category (competition) and overlap among the sites across different categories (mutualism). As a dependent variable, we calculated the number of increased unique visitors between June and October for each site.

We present four major arguments in this paper. First, homophily principle among different categories of people off-line holds true on-line. The Internet users' patterns of on-line interaction and navigating behavior reflect their interests and tastes, which again are strongly affected by demographic attributes. This on-line homophily turns the Internet into a socially constructed space embedded in social relations of its users (cf. Wellman and Gulia, 1999) rather than a social vacuum devoid of social relations and structures. In short, our on-line actions are not separated from off-line social activities but linked with them.

Second, Internet users collectively constitute important environment for the Internet sites. As their attention and interest become scarce resources, we can think of them an important dimension of Internet site niche. Internet users' homophily based on demographic attributes bring social structure into the niche, which we can measure with site visitors' demographic variables.

We have put these arguments to empirical test examining the duality between Internet sites and their visitors. We have measured the niche width of Internet sites with the demographic distribution of their visitors. Multidimensional scaling is then employed to locate Internet sites on two dimensional space using the measures of pairwise overlap in site visitors' demographic attributes. On the graphs, we could find meaningful clustering of Internet sites supporting our arguments for on-line homophily and visitor-based Internet niche formation.

Third, Internet sites' niche defined in terms of visitor attributes has significant effects on the site performance. Overlap of visitor niche presents two possibilities to Internet sites. On one hand, there is a possibility of mutualism, in which the sites benefit from synergy and cooperation among the sites sharing the niche. On the other hand, niche overlap could be a liability to the sites facing pressures of exclusionary competition. The outcome depends on whether the Internet sites sharing the same niche are complementary or substitutable in their contents and function. Within each category, we claim, substitutability would prevail and put competition into work. Across different categories complementarity would dominate and activate mutualism. Our results from regression models of site performance strongly suggest the benefits of mutualism but not of competition.

Lastly, strategies of Internet sites interact with their niche and jointly affect site performance. The distinction of generalism and specialism relates to the niche width of Internet sites. As popular expression holds, inter-site distances are just a click away, ample possibilities exist for Internet sites to take advantage of inter-relatedness and one way to realize that possibility is to become a portal site. Internet sites can get the most out of mutualism by becoming a portal. Estimating separate regression models for portal and non-portal revealed beneficial effect of niche overlap for portal sites but no such effect for non-portal. These results serve as empirical evidences for our fourth argument.