

Direct Network Effects, Small World Networks, and Industry Formation

by

Jeffrey L. Funk

Associate Professor

National University of Singapore

etmfjl@nus.edu.sg

Forthcoming

Telecommunications Policy

Direct Network Effects, Small World Networks, and Industry Formation

Abstract

This paper addresses the formation of industries that involve direct network effects. Using two concepts from the literature on network effects (critical mass and inverse demand curves) and descriptive data from the formation of five telecommunication-related industries, this paper argues that a critical mass of users was created multiple times in these industries where multiple critical masses of users can be represented as local maximums in an inverse demand curve. The existence of these multiple local maximums reflects the existence of different sub-populations of users within a total population of potential users where these populations of users can be considered small world networks. Initially the different sub-populations represent fragmented networks of users that are served by fragmented networks of firms. Over time connected networks of both firms and users emerge.

1. Introduction

This paper links two insufficiently answered and seemingly unrelated questions. First, how do new industries emerge and their associated products diffuse (Rogers, 1962; Rostow, 1991) when there are strong direct network effects? The existence of strong direct network effects can require the creation of a “critical mass of users” before growth will occur (Arthur, 1994; Shapiro and Varian, 1999) where a critical mass of users can be represented using an inverse demand curve (Economides and Himmelberg, 1995). According to Rohlfs (1974) seminal paper and subsequent research on critical mass, a user’s willingness to pay is a function of quantity in an inverse demand curve (as opposed to *visa versa* in a traditional demand curve) where the willingness to pay initially rises as the quantity (i.e., number of users) rises due to the existence of strong network effects. The number of users must reach a certain level (i.e., a critical mass of users) in order for growth to continue (Rohlfs, 2001; Economides and Himmelberg, 1995). But how does a critical mass of users emerge, do they emerge in a single population or in multiple sub-populations and if they emerge in multiple sub-populations are these sub-populations later connected and if so how? For example, did the early growth in the telephone industry occur in a single or multiple sub-populations of users and if the latter is the case, how were these sub-populations of users later connected?

Second, what role do social networks play during industry formation? Scholars have applied network theory to firms (Kogut, 2000; Dyer and Nobeoka, 2000), individuals (Burt, 1997; Coleman, 1990), and physical and chemical phenomena (Strogatz, 2003). Most connected networks appear to display small-world characteristics of relatively short path length and high clustering. In social networks path length refers to the

number of connections that separate people in a population and clustering refers to the extent to which these connections are interdependent and thus primarily exist within small groups of people (i.e., within sub-populations) (Milgram, 1967; Uzzi, 1996; Kogut and Walker, 2001; Davis et al, 2003). Watts and Strogatz's seminal research (Watts and Strogatz, 1998; Strogatz, 2003; Watts, 2003) suggests that these small-world networks strike a balance between ordered networks (ones with long path lengths and high clustering) and random networks (ones with short path length and low clustering). But where do these small world networks come from? Do they come from these ordered or random networks or perhaps from the fragmented networks referred to in the previous paragraph (Watts, 2003; Baum et al, 2003)?

This paper combines the concepts of critical mass, inverse demand curves, and small-world-network theory to present a conceptual framework for addressing these questions. The framework builds on existing theoretical work on critical mass and inverse demand curves (Rohlf's, 1974, 2001; Economides and Himmelberg, 1995) and uses descriptive data on five industries to show that 1.) a critical mass of users must be created multiple times in some industries; and 2.) these multiple critical masses of users can be represented as multiple "bumps" in an inverse demand curve.

. Drawing on small world network theory (Watts and Strogatz, 1998; Strogatz, 2003; Watts, 2003), the framework shows how 3.) the existence of these multiple bumps reflect the existence of multiple sub-populations within a potential population of users; 4.) these sub-populations initially represent fragmented networks of users that are served by fragmented physical networks and fragmented networks of firms; and 5.) the interaction between firms and users may lead overtime to the emergence of connected networks of both firms and users where these networks display small-world

characteristics.. Sixth, combining the concepts of small world networks and network effects, the framework shows how the initial growth in the industry occurred primarily through increases in the number of fragmented networks and not through growth in one network. Furthermore, the growth in these fragmented networks delayed the emergence of a “dominant” network and sometimes prevented the first network from becoming the dominant one. Although some research has hypothesized the existence of multiple bumps in an inverse demand curve (Rohfs, 2001), to our knowledge no one has attempted to place names on these bumps and combined items 1 and 2 with items 3-6.

This paper first discusses the existing literature on critical mass, inverse demand curves, and small-world networks that are relevant to the proposed framework. Second, it applies the framework to the formation of five industries that exhibit direct network effects and that required a critical mass of users to be created multiple times. It focuses on the formation of these industries in the U.S. since the U.S.’s institutional characteristics (Kogut, 2000) (e.g., low regulatory barriers to entry) have enabled a larger diversity of firms to participate in industry formation than in other countries (Nelson, 1993; Mowery and Rosenberg, 1998); this diversity of firms and thus product offerings has made it easier to identify the sub-populations that are represented by the multiple bumps in an inverse demand curve.

2. Key concepts/proposed framework

Inverse demand curves are often used to represent the demand for products that display strong network effects (Rohlf, 1974, 2001). They plot price (willingness to pay) as a function of quantity (as opposed to quantity as a function of price in a traditional demand curve). The willingness to pay rises as the quantity (e.g., the number of users)

rises when: 1) there is zero utility in a network of zero size; or 2) there are immediate and large external benefits to the expansion of very small networks (Economides and Himmelberg, 1995). This greater willingness to pay reflects the existence of strong network effects (Arthur, 1994; Katz and Shapiro, 1985, 1986, 1994). This paper focuses on those products that exhibit direct network effects (e.g., telephone) and there is zero utility in a network of zero size, i.e., zero utility for the first user (Economides and Himmelberg, 1995).

Because there is zero utility in a network of zero size, a critical mass of users must be created in these industries in order for growth to continue. The critical mass of users is defined as the number of users on the left side of the inverse demand curve that correspond to each price (See Figure 1). Since the left side of the curve is unstable, the achievement of a critical mass of users causes the number of users to rise to the level corresponding to the right hand side of the curve (Rohlfs, 2001; Economides and Himmelberg, 1995). Without a critical mass of users, the number of users will return to zero. This occurred with AT&T's Picture Phone Service (described in the results section) and many products that display indirect network effects such as digital audio tape, digital compact cassette, mini-discs, high-definition television, (Rohlfs, 2001; Grindley, 1995), and AM stereo (Shapiro and Varian, 1999).

Place Figure 1 about here

Another way to emphasize the importance of a critical mass of users is to describe how supply curves interact with demand curves and how this interaction is different when a critical mass of users is required. First, demand and supply curves are

independent constructs; although changes in supply do impact on the point at which the demand and supply curves intersect and thus on the demand for products and services, changes in supply do not impact on the *shape* of the demand curve or cause movements in it (Samuelson and Nordhaus, 2005). Second, although demand curves may move over time for other reasons (e.g., through changes in income or through the introduction of complementary or competing products), supply curves generally move much more than demand curves do through changes in technology (Samuelson and Nordhaus, 2005) such as in the manner shown in Figure 2. The supply curves for the cases addressed in this paper have all experienced large movements to the right as shown in Figure 2 as evidenced by large reductions in the prices of telephone services (Brock, 1981), wireless services (Garrard, 1999), facsimile machines (Peterson, 1995), and Internet-related services (Abatte, 1999; Kenney, 2003) such as mail and video conferencing ones. When there is zero utility in a network of zero size, the supply curve must move further to the right than would be needed in a product that has a positive utility in a network of zero size since price-insensitive consumers do not exist until a critical mass of users has been created. This may delay the start of growth until technological improvements move the supply curve further to the right than would be needed in a product that has a positive utility in a network of zero size (Rohlf's, 2001).

Focusing on the concept of an inverse demand curve, some research suggests that there are multiple “bumps” (i.e., local maximums) in such a curve (See Figure 3 for an idealized version) where a critical mass of users must be created in each bump of the inverse demand curve (Rohlf's, 2001). Each bump represents a sub-population of users that can also be thought of as a fragmented network of users within the entire population of potential users. The bumps in the left hand side of Figure 3 represent users with a

greater willingness to pay than those represented by the bumps in the right hand side of the figure. A mathematical appendix in Rohlfs (2001) shows a figure of an inverse demand curve that has multiple bumps and the cases in the main part of his book *imply* that multiple bumps exist in the demand curves for many industries, albeit Rohlfs does not actually reference this figure in the cases in the main part of the book.

Place Figures 2 and 3 about here

The existence of these different sub-populations enables multiple fragmented networks to emerge in an industry that has strong network effects. Without the existence of multiple sub-populations, the literature suggests that network effects would cause growth to primarily occur within those networks that first experienced growth (Arthur, 1994; Katz and Shapiro, 1985, 1986; 1994). However, the existence of multiple sub-populations can cause the initial growth in the industry to occur more through the creation of new fragmented networks, each serving different sub-populations of users, than through the growth inside individual networks.

The existence of multiple sub-populations and fragmented networks are consistent with some descriptions of how small world networks emerge. Small-world networks exhibit short path lengths and high clustering. Path length refers to the number of connections that separate people in a population and clustering refers to the extent to which these connections are interdependent and thus primarily exist within small groups of people (i.e., within sub-populations) (Milgram, 1967; Uzzi, 1996; Kogut and Walker, 2001; Davis et al, 2003).

Building from this research and from his seminal paper with Strogatz (Watts and

Strogatz, 1998), Watts (2003) draws analogies between ordered networks and so-called “caveman worlds” and between random networks and so-called “Solaria” networks where small-world networks still occupy a position between ordered and random networks. He does this by using the variable alpha to characterize the rules of interaction between individuals (See Figure 4). When individuals only interact with those they know, alpha is zero; when the interactions are random, alpha is infinity. Low values of alpha lead to so-called “caveman worlds” that sound similar to the sub-populations and fragmented networks that are mentioned above. In these caveman worlds, individuals only interact with a small number of people within a sub-population (Watts uses the term cluster) and thus there are short path lengths within these small clusters (i.e., sub-populations). As alpha increases, individuals interact with people outside their sub-population (i.e., outside their cluster), thus causing the path length to increase until all the individuals in a population are connected in one network. Further increases in alpha cause the path length (and later the clustering coefficient) to fall perhaps leading to small-world networks or even a Solaria-network (Watts, 2003) where the interactions between people are random.

Place Figure 4 about here

This change from the fragmented networks of the caveman world to a connected network with small-world characteristics or a Solaria-type one (Watts, 2003) is partly driven by the emergence of physical connections between fragmented networks. These connections may result from the internal growth, acquisitions, or alliances of firms (Harrigan, 1985; Kogut, 1988; Jarillo, 1988; Dyer and Nobeoka, 2000) and/or the

creation of standards. These connections may cause a network of firms to emerge that reflect the institutional arrangements and the inherent characteristics of the technologies that populate the industry (Kogut, 2000). In the case of standards, the creation of them may enable specific firms to dominate these industries through their control of them or the processes used to set them (Katz and Shapiro, 1985, 1986, 1994; Shapiro and Varian, 1999). On the other hand, depending on the circumstances (which are not covered in this paper), the emergence of fully open standards can cause competition to eventually shift from “between bumps” to within specific modules within the physical system/network thus leading to vertical disintegration in the industry where the physical system serves most or all of the population of users (Langlois and Robertson, 1992; Baldwin and Clark, 2000; Brusoni and Prencipe, 2001; Langlois, 2003; Chesbrough, 2003).

3. Methodology

The author first conducted a preliminary analysis of more than 15 industries that display direct (fixed line telephone, wireless, facsimile, Internet mail, video telephones/conferencing, short messaging services, telegraph, modems, social networking sites, instant messaging) and indirect (recorded music, radio and television broadcasting, personal computers, video, video games, world wide web, mobile Internet) network effects. This analysis considered academic literature from history, sociology, business, and economics and journalistic accounts of these industries. It revealed that the concepts of critical mass, inverse demand curve, and small-world networks apply differently to industries that exhibit direct and indirect network effects. For example, the variable “quantity” in an inverse demand curve can be more easily

interpreted as the number of users in industries that exhibit direct effects than in ones that exhibit indirect network effects and for this reason the paper only focuses on industries that exhibit direct network effects.

For industries that exhibit direct network effects, industries were eliminated from the analysis for other reasons. Sufficient information was not found on short messaging services, modems, social networking sites, and instant messaging services while other industries/products exhibited positive utility in a network of zero size (i.e., for the first user). For example, although the telegraph displayed direct network effects, a critical mass of users was not needed for growth to occur since the first telegraph line provided value to users (Brock, 1981) and would have continued to provide users with value even if the number of users or telegraph lines had not increased.

A more detailed analysis was then conducted of the five selected industries/products that display direct network effects. Since other scholars have used data on price and quantity to quantitatively demonstrate the role of a critical mass of users in some of these industries such as telephones (Rofls, 1974) and facsimile machines (Economides and Himmelberg, 1995) and many other scholars have shown how the prices of telephones (Brock, 1981), wireless services (Garrard, 1999), facsimiles (Peterson, 1995), and Internet-related services (Abatte, 1999; Kenney, 2003) such as mail and video conferencing have dropped over the years through changes in technology and thus caused movements in the supply curve, this paper focuses on identifying the bumps in the inverse demand curves, their associated sub-populations, and the process of connecting these bumps for both firms and individuals in these industries. The literatures mentioned in the previous paragraph were used to do this where academic histories and journalistic accounts, particularly older ones, were emphasized in order to

obtain “raw” data on the early days of the industries.

The bumps in the inverse demand curves were defined by identifying distinct sub-populations of users that emerged through the growth in the number of fragmented networks (as opposed to the growth inside individual networks). Although growth in an individual network does not preclude the existence of multiple sub-populations and bumps in an inverse demand curve, this paper’s analysis made the existence of multiple fragmented networks a prerequisite for defining multiple bumps in an inverse demand curve for a specific industry. The ordering of the bumps from left to right in an inverse demand curve was done by identifying the order in which the sub-populations adopted the products that emerged with these industries. The process by which these bumps (i.e., fragmented networks) were connected over time was also done by identifying the way in which sub-populations of users were connected over time and how this impacted on competition.

4. Results

Table 1 summarizes the bumps in the inverse demand curve for five industries that display direct network effects and that required the creation of a critical mass of users before growth could occur. The order of bumps listed in Table 1 corresponds to the order of bumps (from left to right) in an inverse demand curve (See Figure 3).

Place Table 1 about here

4.1 Telephones

Bell Telephone and Western Union’s telephone subsidiary (American Speaking

Telephone) began independently leasing telephones in 1877 in the U.S. to individual firms. Although they managed to lease more than 3000 phones (and lines connecting them) to brokerages, newspaper offices, hotels, railways, and other large corporate users of telegraph services (Coon, 1939; Huurdeman, 2003), it quickly became apparent that connecting phones to a switching center would provide lower costs and provide greater benefits to users. Both Bell Telephone and Western Union began to independently franchise local operating companies in major cities beginning in February 1878. After Bell Telephone's patent infringement suit caused Western Union to exit the business and sell its network of 56,000 phones to Bell Telephone in 1879 (Brock, 1981), Bell Telephone basically had a monopoly until its patents expired in 1894 (Brock, 1981; Mueller, 1997; Rohlf, 2001).

The expiration of Bell's patents in 1894 led to the entry of new firms and their start of commercial services in cities and the entry of "mutual systems" and "farmer lines" in rural areas. For the commercial services the new entrants targeted lower income groups in cities many of which displayed little interest in communicating with higher income groups (Brock, 1981; Mueller, 1997). Merchants and farmers in rural areas implemented "mutual companies" and "farmer lines" respectively using their own capital. By 1902 there were more than 6000 of them in the U.S. (Fischer, 1987) and 222 mutual companies were started between 1900 and 1917 just in Southeastern Iowa, or more than eight times the number of commercial companies there (Barnet and Carroll, 1987). The farmer lines often used fences as wires and did not include switchboards (Fischer, 1987). The small number of average subscribers for these farmer lines (11), mutual systems (90), and even non-Bell commercial systems (700) in 1902 (Brock, 1981; Sterling et al, 2006) highlights the fragmented nature of the U.S. telephone industry in 1902. The large

number of these commercial systems, mutual companies and farmer lines, which did not emerge in most of the “centrally” controlled European countries, is considered a major reason why the penetration rate of telephones in the U.S. had exceeded that of most European countries by 1920 (Fischer and Carroll, 1988; Brock, 1981).

The non-Bell commercial systems, mutual companies, and farmer lines are often called independents in order to differentiate them with Bell Telephone. It was not until connections began to be made between these independents and between them and Bell Telephone (i.e., connections between fragmented networks) that long-distance services and thus network effects became a large advantage for Bell Telephone (it became AT&T in 1899). Until then the growth in subscribers occurred more through the new entrants’ creation of new fragmented networks that served different sub-populations of users than through the growth inside individual networks. But in response to the growth in the number of these fragmented networks (i.e., number of independents) and the resulting drop in AT&T’s share, Theodore Vail introduced a new strategy in which AT&T began using acquisitions and alliances in the early 1900s to provide users with more access to long-distance services than the other telephone companies could do. In particular, AT&T provided its users with not only long-distance access through its local affiliates but also through alliances with companies that did not compete with an AT&T local affiliate (Brock, 1981; Mueller, 1997; Rohlfs, 2001; Sterling et al, 2006).

This description suggests that the demand curve for telephone services can be represented by many narrow, tall bumps and a smaller number of small, wide bumps where each bump represents a sub-population of users and a critical mass of users had to be created in each of these sub-populations. The narrow, tall bumps represented the demand for intra-firm systems by brokerages, newspapers, railways, and other firms and

the demand for farmer lines. The wider bumps represented the demand for mutual companies and commercial services where the incorporation of the intra-firm systems into city-wide commercial services suggests that some of the narrow, tall bumps were also inside some of the small, wide bumps that represented the city wide-commercial services.

The connections between intra-firm systems and city-wide commercial services represent the first connections of fragmented networks. The connections between competing services in a single city and the later long distance connections between cities represent further connections between fragmented networks. Using Watts (2003) terminology, the early telephone represented a caveman world where individuals could primarily communicate only with those people whom their superiors thought they should be communicating with (intra-firm applications) or with people in their own social class (early commercial systems). As the use of commercial telephone services spread, individuals began to converse with a broader range of people over the telephone thus causing a small-world network to emerge; this small world network sometimes displays characteristics of “Solaria” in which telemarketers and other people randomly bombard us with sales pitches and surveys (Collins, 1979).

Making these connections between fragmented networks also reflects a change in competition from competition “within bumps” to competition “between bumps” of which AT&T’s strategic use of long distance services is the largest example. AT&T’s increasing market share led to the Willis-Graham Act of 1921 that made Bell a regulated monopoly and reinforced Bell’s position at the apex of a network of service providers and equipment suppliers in the U.S. telephone industry (Mueller, 1997; Brock, 1981). It was not until competition and interoperability was reintroduced in the 1980s that

AT&T's control over the network of firms in the industry and the competition "between bumps" changed to competition within independent modules/vertically disintegrated layers (Brock, 1994; Sterling et al, 2006).

4.2 Wireless (radio and mobile phone)

Radio was initially introduced as a form of wireless telegraph at the turn of the 20th century. Firms leased this equipment from equipment suppliers such as Marconi where shipping companies were major users because "ship-to ship" and "ship-to shore" communication with traditional telegraph lines was difficult to do for physical reasons. Improvements in transmitters and receivers enabled wireless voice communication and led to the emergence of a second set of users in the early 1910s (Douglas, 1987). So-called "radio heads" began using ham radio to converse semi-anonymously over long distances, which led to various laws including the Radio Act of 1912 (Lewis, 1991; Spar, 2001).

Following the explosive growth in commercial radio broadcasting in the early 1920s and the reduction in the cost of receivers that accompanied this growth (Sterling, 1979; Lewis, 1991; Sobel, 1986), governments and private organizations began implementing so-called "private mobile radio" systems. The Detroit Police Department is often considered the first organization to implement such a system in 1921. Subsequently, tens of thousands of government (police, fire, forestry conservation, highway maintenance, local government services, military) and private (power, oil, motion picture, telephone maintenance, transportation, taxi, trucking) organizations introduced their own proprietary systems in the U.S. (Yacoub, 1993; Coe, 1996; Garrard, 1999). These systems are completely independent from public systems that were first

introduced in the 1940s by AT&T. AT&T's service used a single transmitter to provide services to a small number of people in a very broad area and it was largely replaced by so-called cellular systems in the early 1980s (Coe, 1996; Garrard, 1999).

This description suggests that the demand for wireless services can be represented by at least three large bumps (wireless telegraph users, ham radio users, private mobile radio users). The first and third bumps each consist of many smaller bumps where each of the smaller bumps represent a specific firm's demand for leasing for example Marconi's wireless telegraph equipment (first large bump) or implementing a private mobile radio system (the third large bump). The second bump of ham radio users may also have consisted of many smaller bumps where each small bump may have represented the demand in different geographical areas or in different topics for discussion among the ham radio operators and thus different sub-populations of users. Just in the third bump, the diffusion of private mobile radio systems occurred more through the creation of new fragmented networks that served different sub-populations of users than through increases in the size of individual networks.

The introduction of cellular phone systems in the early 1980s led to the emergence of networks that connected the separate sub-populations and their fragmented networks. Using Watts (2003) terminology, the users of Marconi's early equipment, private mobile radio, and to a lesser extent ham radios lived in a caveman world where they could primarily communicate with small numbers of people in specific sub-populations. Although AT&T did introduce a public system in the 1940s, there were few users and it was the rapid diffusion of cellular phones in the 1980s and 1990s (Garrard, 1999) that led to the emergence of a small-world network; this small world network sometimes displays characteristics of "Solaria" in which we can be interrupted and our privacy

invaded by almost anyone at anytime by calls on our mobile phones.

The introduction of cellular phones also caused competition between equipment providers to change from “within bumps” to “between bumps.” In private mobile radio, manufacturers supplied individual “bumps” such as individual fire and police departments, taxi companies, and military services with proprietary systems that were incompatible with other systems. With the introduction of cellular phones in the 1980s came standards and the move towards a single interconnected system initially at the national and later global levels. Americas’ first generation analog standard, AMPS (Advanced Mobile Phone System), facilitated this in the U.S. while Europe’s digital standard GSM (Global System Mobile) facilitated this at the European and later global levels (Garrard, 1999). Furthermore, the emergence of global standards such as GSM caused a small number of firms such as Nokia, Ericsson, and Motorola to connect people at the global level and to occupy central positions within the standard setting organizations for GSM (Funk, 2002) and thus the overall firm networks in the industry.

4.3 Facsimile

Although the first facsimile machine was built in 1843 and the first service was offered in 1865, facsimile machines did not begin diffusing to any great extent until the 1980s (Rohlf, 2001). For most of the 20th century, firms used facsimile machines for the intra-firm exchange of pictures and other information that could not be easily transmitted via telex (the telex was the 20th century version of the telegraph). This information included pictures for newspapers, weather maps for newspapers and other firms, fingerprints and mug shots for police, and other information for internal communications within railroads and other large organizations (Peterson, 1995;

Costigan, 1971). Between 1945 and 1970, about 50,000 machines were sold in Japan (by one firm) (Business Week, 1970) and about 15,000 in the U.S. where machines from different manufacturers used different standards and thus could not communicate with each other (Peterson, 1995; Rohlfs, 2001).

The use of facsimile machines for inter-firm communication began to increase in the 1970s and 1980s as restrictions on connecting third-party equipment to phones were eliminated, standards were agreed upon, prices for the machines fell, and facsimile transmission services were introduced. Restrictions on connecting third-party equipment to phones were reduced with the Hush-a-Phone ruling in 1956 and eliminated with the Carterfone ruling in 1968 (Brock, 1981). Western European telecommunication providers agreed on standards in 1968 for so-called Group 1 machines and these standards were updated for Group 2 and Group 3 machines in 1976 and 1980 respectively with the participation of Japanese and U.S. firms. By the mid-1980s, facsimile machines had become an essential tool for businesses and some consumers (Peterson, 1995; Rohlfs, 2001). The importance of facsimile machines in the 1980s was reflected in the number of firms offering facsimile transmission services; the number of these businesses just in Manhattan had exceeded 60 by 1989 (Baum and Korn, 1995).

This description suggests that the demand for facsimile machines and services can be represented by at least two wide bumps (1. intra-firm communication; and 2. inter-firm and consumer communication through facsimile transmission services) in which a number of tall and narrow bumps probably existed within the wide bump of intra-firm communication. Although many tall and narrow bumps may also have existed within the other wide bump of inter-firm and consumer communication through facsimile transmission services, the fact that business users, consumers, and facsimile

transmission services all used the same standards and thus the same network of facsimile machines does not allow us to say that fragmented networks and thus many tall and narrow bumps existed within this second wide bump. Nevertheless, like the intra-firm networks for telephones and for private mobile radio systems, the early diffusion of facsimile machines occurred more through the creation of new fragmented networks that served different sub-populations (primarily in firms) than through the growth inside individual networks.

The agreements on standards enabled connections to be made between the fragmented networks. They enabled facsimile machines from different manufacturers to communicate with each other (Peterson, 1995), which enabled different firms and households to exchange faxes. Although it can be said that the emergence of these standards caused the “caveman” world (Watts, 2003) that existed in the early days of facsimile usage to be replaced with a form of small-world network, this small world network probably displayed fewer characteristics of “Solaria” than those of telephones and mobile phones have done.

The emergence of these standards also reflects a change in competition from “between” to “within” bumps. For example, Japanese firms that supported open standards moved faster to connect the bumps than did U.S. firms that focused on proprietary systems and their customers that used them. Their support of open standards and the success this brought also reflected the fact that Japanese firms began to occupy central positions within the network of facsimile manufacturers including central positions within the standard setting organizations (Rohlfs, 2001; Peterson, 1995).

4.4 Internet Mail

Beginning with the Defense Department, different parts of the U.S. government began to independently fund the development and introduction of packet-based systems such as ARPAnet, NSF Net, and Aloha Net in the late 1960s and early 1970s (Abbate, 1999; Segaller, 1998). Computer science departments and research institutes were the first organizations to implement these systems in which mail unexpectedly ended up being the most widely used application. Researchers used the mail for communication within their small circles of colleagues. As universities gradually expanded Internet mail access to other university departments in these fragmented networks in the late 1970s (Abbate, 1999; Mowery and Simcoe, 2002), firms also began to implement intra-firm networks that used packet-based technology and that primarily supported the use of Internet mail within their firms. Early users included Volvo (Astebro, 1995), Manufacturers Hanovers Trust (Nyce and Groppa, 1983), DEC (Crawford, 1982), AT&T, Bank of America, HP, IBM, Westinghouse, Xerox, 3M, and Peat Marwick (Sproull and Kiesler, 1986). The introduction of local area networks (LANs) for connecting personal computers and printers complemented the introduction of these intra-firm mail networks (von Burg, 2001). Telenet was the first firm to offer public packet-switched services in 1975 and they were followed by spin offs from some of the government-sponsored networks such as NSF Net and by startups such as American Online (AOL), CompuServe, and Prodigy in the 1980s (Abbate, 1999; Mowery and Simcoe, 2002; Rohlfs, 2001, Kenney, 2003).

Connections between these fragmented networks began to accelerate in the late 1980s and early 1990 as mergers, acquisitions, and alliances occurred within these providers of commercial mail services, thus turning these services into so-called

“commercial gateways” and also as long-distance telephone companies began implementing communication “backbones” (Abbate, 1999; Segaller, 1998). Direct network effects played an important role in this competition where the growth in users primarily occurred in the largest commercial gateways such as AOL (Shapiro and Varian, 1999; Rohlfs, 2001).

This description suggests that the demand for the initial Internet mail services can be represented by at least three bumps (computer science and other university departments, intra-firm networks, and general business/consumers) in the inverse demand curve where at least the first two wide bumps consist of many narrow and tall bumps. Each bump represented a sub-population of potential Internet users where the members of these sub-populations were primarily interested in communication within their own sub-population. Like the other industries covered in this paper, the diffusion of Internet mail services occurred more through the creation of new fragmented networks than through increases in the size of individual networks.

Mergers, acquisitions, and alliances gradually turned these commercial mail services into commercial gateways that connected the sub-populations represented by the multiple bumps in the inverse demand curve. Like the early telephone, early Internet mail represented a caveman world (Watts, 2003) where individuals could communicate only with those people that also inhabited the rarified world of the Internet. However, the emergence of commercial gateways connected fragmented networks and brought Internet mail to the masses. Individuals began to exchange mail with a broader range of people thus causing small world networks to emerge that sometimes display characteristics of “Solaria” where we are bombarded with spam and viruses.

There have also been large changes in competition as connections between

sub-populations were made. The emergence of commercial gateways changed the competition from “within bumps” to “between bumps” in Internet mail of which the rise of AOL including its merger with Time Warner is the most well-known example. AOL’s large number of subscribers enabled its users to exchange mail with a larger number of people than subscribers to other services. However, the emergence of open standards for mail and related services has made it easy for anyone to communicate “between bumps” (Abbate, 1999; Rohlfs, 2001). This is reflected in the emergence of vertical disintegration in the Internet, competition “within modules,” and the falling fortunes of AOL.

4.5 Video conferencing/telephones

Beginning with a demonstration at the NY Worlds Fair in 1964, AT&T introduced a number of experimental two-way video communication services in the 1960s and 1970s (Dickson, 1973; Egidio, 1988) of which the largest one was in Chicago and was called Picturephone. About 200 people leased the equipment for \$86.50 a month and shortly thereafter AT&T discontinued the service. In a post mortem analysis, the most common complaint given in interviews was there was no one to talk to. Apparently very few subscribers knew each other (Noll, 1992; Rohlfs, 2001).

Instead, critical masses of users were created by firms that first installed video telephones in the form of one-way and later two-way video conferencing systems for intra-firm communication. By 1985, firms such as Aetna Life and Casualty Company, American Hospital, Atlantic Richfield, AT&T, Boeing, Digital Equipment Corporation, Ford, HP, Kodak, McDonnell Douglas, J.C. Penney, TI, and Wang had installed intra-firm systems and a few of them were using the systems for communication with

specific suppliers and customers (Neustadt, 1985). Hotels such as Holiday Inn, the Marriott Corporation, and Hilton Hotels also began to offer such services to their customers in the 1980s (Egido, 1988). The diffusion of PCs, mobile phones, and the Internet and falling telecommunication costs and cheaper cameras has brought video conferencing and video telephone calls to some extent to the masses. So-called desktop video conferencing (Webster, 1998) created new applications such as distance education (Bates, 2005) and health care (Gammon et al, 1996; Huseyin and Iacono, 1999). The diffusion of camera phones has also brought video calls to the mobile phone (Funk, 2004).

This description suggests that the demand for video conferencing/telephone can be represented by at least three bumps (intra-firm networks, inter-firm services, and general business/consumers) in the inverse demand curve where the first bump consists of many narrow and tall bumps. Each bump represented a sub-population of potential users where the members of these sub-populations were primarily interested in communication within their own sub-population. Like the other industries covered in this paper, the diffusion of video conferencing/telephone initially occurred more through the creation of new fragmented networks than through increases in the size of individual networks.

The diffusion of the Internet and the existence of the standards that supported the diffusion of the Internet gradually connected these fragmented networks and the sub-populations they supported. Although it can be said that the emergence of these standards enabled the replacement of the “caveman” world (Watts, 2003) that existed in the early days of video conferencing with some form of small-world network, this small-world network may never display the characteristics of Solaria to the extent that

the telephone, mobile phone, and Internet mail have done. There have also been large changes in competition as connections between sub-populations were made. The emergence of inter-firm systems changed the competition from “within bumps” to “between bumps” where the Internet has provides the basic platform from which these services are offered and has replaced the proprietary systems that were initially used for intra-firm communication (Webster, 1998; Bates, 2005).

5. Discussion

This paper has focused on industries that display strong network effects and it has combined the concepts of critical mass, inverse demand curves, and small world networks to present a conceptual framework of industry formation. This framework builds on existing theoretical work on inverse demand curves (Rohlf, 1974, 2001; Economides and Himmelberg, 1995) to show that multiple local maximums (i.e., bumps) exist in the inverse demand curves for some products/industries that require a critical mass of users before growth can occur. Each bump represents a sub-population of users that can also be thought of as a fragmented network of users within the entire population of potential users.

The existence of these different sub-populations caused growth to initially occur in the analyzed industries more through the creation of new fragmented networks than through growth inside individual networks. Although shifts in the supply curve were necessary for growth to occur, without the existence of multiple sub-populations the network effects would have probably caused growth to primarily occur within those individual networks that were first implemented (Arthur, 1994; Katz and Shapiro, 1985, 1986; 1994). This suggests that individual users were more concerned with the number

of users in their sub-population than with the total number of users in the industry when they considered the adoption of the products considered in this paper.

These results suggests that an analysis of sub-populations and the fragmented networks that serve them can also help us better understand how “tipping” occurs in an industry. The existing literature (e.g., Katz and Shapiro, 1994; Shapiro and Varian, 1999) largely emphasizes the number of users and the advantages for the largest network of users. The analyses of the five industries studied in this paper suggests that the initial growth in the number of fragmented networks can delay this tipping and in some cases delay the tipping until after fragmented systems and sub-populations are connected. Tipping did not occur in any of the industries studied until after the fragmented networks had been connected.

Furthermore, it was only in telephones and Internet mail that the first network ended up becoming the dominant one in the industry. None of the initially fragmented networks in wireless telegraphs, facsimile machines, and proprietary video conferencing systems became the dominant networks in their industries. Instead, new systems and standards that supported these systems (shown in parentheses) became the dominant network in wireless (cellular phone systems and standards that supported them), facsimile machines (second and third generation standards), and video conferencing systems (Internet standards).

Addressing small world network theory (Watts, 2003), the results suggest that when the initial diffusion of a product occurs within a sub-population of users, small world networks of users often emerge from fragmented ones. The initial users live in a kind of caveman world where they are only connected to members of their own sub-population. As the fragmented physical networks became connected, small-world networks emerge

that sometimes display the characteristics of a “Solaria” in which interactions between users become more random. We can say that the value of the variable alpha, which characterizes the interaction between individuals (Watts, 2003), increases as the product diffuses.

Similar conclusions can also be drawn with respect to firm networks. The descriptions of the telephone, wireless, facsimile and Internet suggest that in some cases small-world networks of firms probably emerge from fragmented ones. AT&T, mobile phone and facsimile manufacturers, and Internet mail providers connected the fragmented networks that had emerged from the initial diffusion of the telephone, wireless, facsimile machines, and the Internet. While more research is needed, inverse demand curves and multiple local maximums may play key roles in the emergence of small-world networks of firms.

This framework has at least four implications for firms and policy makers. First, there are not a specific number of users that constitute a critical mass of users. Instead it was the number of users that could provide an acceptable and stable level of utility from the new product, which constituted a critical mass of users. Although hard data was only available in some industries, the popularity of intra-firm communication as a first application suggests that a small number of users constituted a critical mass of users in many of the industries. This was particularly evident in the telephone industry where the average size of for example farmer lines was only 11 users in 1902.

Second, the formation of new industries involves the creation of a critical mass of users multiple times. Firms and policy makers must provide a slightly different solution for each bump (i.e., sub-population of users) in order to create a critical mass of users in each bump. For example, although intra-firm networks represented some of the first

bumps for the telephone, radio/wireless, facsimile, Internet mail, and video conferencing, public and inter-firm networks represented solutions for other bumps in the inverse demand curve. The existence of these multiple bumps is probably one reason why the greater competition in the U.S. market enabled the telephone, wireless applications (at least initially), and Internet mail to diffuse much more rapidly in the U.S. than in Europe. Competition and the decentralization that it brought about in the U.S. increased the number of firms that were looking for these different bumps in the inverse demand curve, the solutions these sub-populations of users were demanding, and thus the chances the appropriate solutions would be found.

Third, firms and governments have used a variety of techniques to find and bridge these bumps. Bell Telephone leased its technology to other commercial providers and this enabled these providers to find new sub-populations of users. Bell Telephone also made alliances with other providers in its promotion of long-distance services that resulted in connecting the bumps represented by different sub-populations. U.S. and European governments promoted open standards for mobile phones and facsimile machines. Governments required telephone companies to allow facsimile machines and other third party devices to be connected to telephone lines. The U.S. government funded the introduction of the Internet in universities and other research organizations.

Fourth, there was in general an evolution of competition from “within bumps” to “between bumps” in all of the industries that experienced growth as it became important to connect the fragmented networks of users. Firms used internal growth and acquisitions to connect these fragmented networks in the telephone and Internet where specific firms such as AT&T and AOL were able to temporarily (very temporarily in the case of AOL) dominate their industries by connecting the fragmented networks. The

emergence of interoperability standards in the 1980s and late 1990s eventually ended AT&T's and AOL's respective control over these firm networks respectively. In wireless and facsimile machines, it was the emergence of standards that enabled connections to be made between the fragmented networks. Firms such as Ericsson, Nokia, Motorola, and later Qualcomm have benefited from their roles in the standard setting organizations for mobile phones while Japanese firms were temporarily the beneficiaries of their roles in the standard setting organizations for facsimile machines.

6. References

- Abbate, J. (1999). *Inventing the Internet*, Cambridge, MA: MIT Press.
- Arthur, B. (1994). *Increasing Returns and Path Dependence in the Economy*, Ann Arbor: University of Michigan Press.
- Astebro, T. (1995). "The Effect of Management and Social Interaction on the Intra-Firm Diffusion of Electronic Mail Systems," *IEEE Transactions on Engineering Management* 42(4): 319-331.
- Barnet, W. and Carroll, G. (1987). "Competition and mutualism among early telephone companies," *Administrative Science Quarterly* 38: 51-73.
- Baldwin, C. and Clark, K. (2000). *Design Rules, Volume 1: The Power of Modularity*, Cambridge: MIT Press.
- Bates, A. (2005). *Technology, e-learning and distance education*, London: Routledge.
- Baum, J. and Korn, H. (1995). "Dominant Designs and Population Dynamics in Telecommunication Services: Founding and Failure of Facsimile Transmission Service Organizations," 1965-1992, *Social Science Research* 24: 97-135.
- Baum, J., Shipilov, A, Rowley, T. (2003). "Where do Small-Worlds Come From?" *Industrial and Corporate Change* 12(4): 697-725.
- Brock, G. (1981). *The Telecommunications Industry: The Dynamics of Market Structure*, Cambridge: Harvard University Press.
- Brock G. (1994). *Telecommunication Policy for the Information Age: From Monopoly to Competition*, Cambridge: Harvard University Press.
- Brusoni, S. and Prencipe, A., (2001). "Unpacking the Black Box of Modularity: Technologies, Products and Organizations," *Industrial and Corporate Change* 10(1): 179-205.

- Burt, R. (1997). "The contingent value of social capital," *Administrative Science Quarterly* 42: 339-365. .
- Business Week (1970). "Moving Images a Japanese Way," August 29: 30.
- Chesbrough, H. (2003). "Towards a Dynamics of Modularity: a cyclical model of technical advance," In *The Business of Systems Integration*, Prencipe, A., Davies, A. and Hobday, M. (Eds), NY: Oxford University Press.
- Coe, L. (1996). *Wireless Radio: A History*, Jefferson, NC: McFarland and Co.
- Coleman, J. (1990). *Foundations of Social Theory*, Cambridge, MA: Harvard University Press.
- Collins, R. (1979). *The Credential Society*, NY: Academic Press.
- Coon, H. (1939). *American Tel & Tel: The Story of a Great Monopoly*, Freeport, NY: Books for Libraries Press.
- Costigan, D. (1971). *Fax: the principles and practice of facsimile communication*, Philadelphia: Chilton.
- Crawford, A. (1982). "Corporate Electronic Mail – A Communication Intensive Application of Information Technology," *MIS Quarterly* 6(3): 1-14.
- Davis, G., Yoo, M. and Baker, W. (2003). "The small world of the corporate elite, 1982-2001," *Strategic Organization* 1: 301-326.
- Dickson, E. (1973). *The Video Telephone*, NY: Praeger.
- Douglas, S. (1987). *Inventing American Broadcasting, 1899-1922*, Baltimore: Johns Hopkins University Press.
- Dyer, J. and Nobeoka, K. (2000). "Creating and Managing a High-Performance knowledge-sharing: the Toyota case," *Strategic Management Journal* 21: 345-367.
- Economides, N. and Himmelberg, C. (1995). "Critical Mass and Network Evolution in

- Telecommunications,” in *Towards a Competitive Telecommunications Industry: Selected Papers from the 1994 Telecommunications Policy Research Conference*, Brock, G. (ed).
- Egido, C. (1988). “Videoconferencing as a Technology to Support Group Work: A Review of its Failure,” *Proceedings of the ACM conference on Computer-supported cooperative work*, Portland Oregon: 13 – 24.
- Fischer, C. (1987). “The Revolution in Rural Telephony, 1900-1920,” *Journal of social history* 21: 5-26.
- Fischer, C. and Carroll, G. (1988). “Telephone and Automobile Diffusion in the United States, 1902-1937,” *The American Journal of Sociology* 93(5): 1153-1178.
- Funk J. (2002). *Global Competition between and within Standards: the case of mobile phones*, London: Palgrave.
- Funk, J. (2004). *Mobile Disruption: The Technologies and Applications Driving the Mobile Internet*, NY: John Wiley & Sons.
- Future Living (2007). <http://davidszondy.com/future/Living/picturephone.htm>. accessed on April 18, 2007.
- Gammon, D., Bergvik, S., Bergmo, T., Pedersen, S. (1996). “Videoconferencing in psychiatry: a survey of use in northern Norway,” *Journal of Telemedicine and Telecare* 2(4):192-198(7)
- Garrard, G. (1998). *Cellular Communications: Global Market Development*, Boston and London: Artech House.
- Grindley, P. (1995). *Standards Strategy and Policy: Cases and Stories*, Oxford: Oxford University Press.
- Harrigan, K. (1985). *Strategies for Joint Ventures*. Lexington, MA: Lexington Books.

- Hüseyin, T. and Iacono, S. (1999). "Diffusion of Telemedicine: A Knowledge Barrier Perspective," *Telemedicine Journal* 5(3): 223 -244
- Huurdeman, A. (2003). *The Worldwide History of Telecommunication Services*, NY: John Wiley.
- Jarillo, J. (1988). "On strategic networks," *Strategic Management Journal* 9(1): 193-210.
- Katz, M. and Shapiro, C., (1985). Network Externalities, Competition, and Compatibility, *American Economic Review* 75(3), pp. 424-440.
- Katz, M. and Shapiro, C. (1986). "Technology Adoption in the Presence of Network Externalities," *The Journal of Political Economy* 94(4): 822-841
- Katz, M. and Shapiro, C. (1994). Systems Competition and Network Effects, *The Journal of Economic Perspectives*, 8(2), pp. 93-115.
- Kenney, M (2003). "The Growth and Development of the Internet in the United States." In Kogut, B (Ed.), *The Global Internet Economy*, Cambridge: MIT Press.
- Kogut, B. (1988). "Joint ventures: Theoretical and empirical perspectives," *Strategic Management Journal* 9(4): 319-332.
- Kogut, B. (2000). "The network as knowledge: generative rules and the emergence of structure," *Strategic Management Journal* 21: 405-425.
- Kogut, B. and Walker, G., (2001). "The small world of Germany and the durability of national networks," *American Sociological Review* 66: 317-335.
- Langlois, R. (2003). "The vanishing hand: the changing dynamics of industrial capitalism" *Industrial and Corporate Change* 12(2): 351-385.
- Langlois, R. and Robertson, P. (1992). "Networks and innovation in a modular system: lessons from the microcomputer and stereo component industries," *Research Policy*

21: 297-313.

Lewis, T. (1991). *The Empire of the Air: the Men who made Radio*, NY: Harper Collins.

Milgram, S. (1967). "The small-world problem," *Psychology Today* 1: 62-67.

Mowery, D. and Rosenberg, N. (1998). *Paths of Innovation*, NY: Cambridge University Press.

Mowery, D. and Simcoe, T. (2002). "Is the Internet a US invention? – an economic and technological history of computer networking," *Research Policy* 31: 1369-1387.

Mueller, M. (1997). *Universal Service, Competition, Interconnecting, and Monopoly in the Making of the American Telephone System*, Cambridge: MIT Press.

Nelson, R., (ed) (1993). *National innovation systems: a comparative analysis*, Oxford: Oxford University Press.

Neustadt, D. (1985). "Action! Camera! It's Time for a Video Meeting," NY Times, November 10.

Noll, A. (1992). "Anatomy of a Failure: Picturephone Revisited," *Telecommunications Policy*, May/June 17(3): 307-316.

Nyce, H. and Groppa, R. (1983). "Electronic Mail at MHT," *Management Technology* 32(11): 65-72.

Peterson, M. (1995). "The emergence of a mass market for fax machines," *Technology in Society* 17(4): 469-482.

Rogers, E. (2002). *Diffusion of Innovations*, NY: Free Press.

Rohlf, J. (1974). "A Theory of Interdependent Demand for a Communication Service," *Bell Journal of Economics* 5 (1): 15-37.

Rohlf, J. (2001). *Bandwagon Effects in High-Technology Industries*, Cambridge, MA: MIT Press.

- Rostow, W. (1991). *The Stages of Economic Growth: A Non-Communist Manifesto*, Cambridge, England: Cambridge University Press.
- Samuelson, P. and Nordhaus, W. (2005). *Economics*, NY: McGraw Hill.
- Segaller, S. 1998. *Nerds: A Brief History of the Internet*, NY: TV Books.
- Shapiro, C. and Varian, H. (1999). *Information Rules*, Boston: Harvard Business School Press.
- Sobel, R. (1986). *RCA*, Briarcliff Manor, NY: Stein and Day.
- Spar, D. (2001). *Ruling the Waves*, NY: Harcourt.
- Sproull, L. and Kiesler, S. (1986). "Reducing social context cues: electronic mail in organizational communication," *Management Science* 32(11): 1492-1512.
- Sterling, C. (1979). "Television and Radio Broadcasting," in *Who Owns the Media?* Compaine, B. (Ed): 61-126. White Plains, NY: Knowledge Industry Publishers.
- Sterling, C., Bernt, P., and Weiss, M. (2006). *Shaping American Telecommunications: A History of Technology, Policy, and Economics*, Mahwah, NJ: Lawrence Erlbaum Associates.
- Strogatz, S. (2003). *Sync: How Order Emerges from Chaos in the Universe, Nature, and Daily Life*, NY: Theia.
- Uzzi, B. (1996). "The sources and consequences of embeddedness for the economic performance of organizations: the network effect," *American Sociological Review* 61: 674-698.
- von Burg U. (2001). *The Triumph of Ethernet*, Stanford: Stanford University Press.
- Watts, D. (2003). *Six Degrees: The Science of a Connected Age*, NY: Norton.
- Watts, D. and Strogatz, S. (1998). "Collective dynamics of 'small world' networks," *Nature* 393: 440-442.

Webster, J. (1998). "Desktop Videoconferencing: Experiences of Complete Users, Wary Users, and Non-Users," *MIS Quarterly* 22(3): 257-286

Yacoub, M. (1993). *Foundations of mobile radio engineering*, Boca Raton, FL: CRC Press.