Telecommunications Policy 33 (2009) 241–252

Contents lists available at ScienceDirect

Telecommunications Policy

URL: www.elsevierbusinessandmanagement.com/locate/telpol

Direct network effects, small-world networks, and industry formation

Jeffrey L. Funk *

National University of Singapore, 7 Engineering Drive 1, Block E3A, #04-07, Singapore

ARTICLE INFO

Keywords:
Network effects
Critical mass
Inverse demand curve
Small-world networks
Telephone
Facsimile
Wireless
Video
Internet mail

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, 2002; 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 & Varian, 1999) where a critical mass of users can be represented using an inverse demand curve (Economides & 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 (Economides & Himmelberg, 1995; Rohlfs, 2001). 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 (Dyer & Nobeoka, 2000; Kogut, 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

* Tel.: +65 6776 6516; fax: +65 6776 0755.
E-mail address: etmfjl@nus.edu.sg

0308-5961/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.telpol.2009.02.003

groups of people (i.e., within sub-populations) (Davis, Yoo, & Baker, 2003; Kogut & Walker, 2001; Milgram, 1967; Uzzi, 1996). Watts and Strogatz's seminal research (Strogatz, 2003; Watts, 2003; Watts & Strogatz, 1998) 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 (Baum, Shipilov, & Rowley, 2003; Watts, 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 (Economides & Himmelberg, 1995; Rohlfs, 1974, 2001) 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 (Strogatz, 2003; Watts, 2003; Watts & Strogatz, 1998), the framework shows how (3) the existence of these multiple bumps reflects 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 theses networks display small-world characteristics. (6) 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 (Rohlfs, 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. Secondly, 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 US since the US'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 (Mowery & Rosenberg, 1998; Nelson, 1993); 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 (Rohlfs, 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 & Himmelberg, 1995). This greater willingness to pay reflects the existence of strong network effects (Arthur, 1994; Katz & 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 & 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 Fig. 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 (Economides & Himmelberg, 1995; Rohlfs, 2001). 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 (Grindley, 1995; Rohlfs, 2001), and AM stereo (Shapiro & Varian, 1999).

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 & 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 & Nordhaus, 2005) such as in the manner shown in Fig. 2. The supply curves for the cases addressed in this paper have all experienced large movements to the right as shown in Fig. 2 as evidenced by large reductions in the prices of telephone services (Brock, 1981), wireless services (Garrard, 1997), facsimile machines (Peterson, 1995), and Internet-related services (Abbate, 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