

5

Paths and Regions- The Creation and Growth of Silicon Valley

Martin Kenney
University of California, Davis
Berkeley Roundtable on the International Economy

Urs von Burg
University of St. Gallen, St. Gallen, Switzerland

For the last 40 years, a geographic and mental space in the San Francisco Bay Area now known as Silicon Valley has been the birthplace of many of the largest and fastest growing electronics firms in the world and a number of new industrial sectors.¹ The technologies these firms commercialized have had a significant impact on many aspects of social and economic life. To facilitate the commercialization, a set of institutions evolved in Silicon Valley to nurture the new firms; these were established to exploit the potential for rapid growth stemming from electronics innovations.

The observation that technologies and places have histories and that these histories matter is, by itself, unremarkable (see Bassanini & Dosi, chap. 2, this volume). If, to invert Voltaire's (1999) Dr. Pangloss, it is accepted that the current situation is not necessarily the best of all possible worlds; then we would leave the world of microeconomics and enter a world of struggle, strategy, and serendipity, in which human beings working alone

¹This region is also the home of a large number of medical equipment firms and the largest concentration of biotechnology firms in the United States. These startups derived critical benefits from proximity to institutions such as venture capital described later in this chapter.

and in groups, create novelty, while being conditioned by their history. In this world, the theoretical models of microeconomics lose power and can only be accepted as partially valid, at best; more often, are irrelevant. Freed from the simplistic, totalizing, ahistorical model of microeconomics, we are regrettably confronted with complexity.

Path dependency directs inquiry toward the ways in which today's realities are based on yesterday's events. Silicon Valley is, in many ways, an ideal case for examining the strengths and weaknesses of path dependent explanations. The concepts from path dependency literature provide a useful departure point for understanding the creation and evolution of Silicon Valley. For one thing, Arthur (1994) specifically referred to Silicon Valley as an example of path dependent industrial clustering due to agglomeration effects. Implicit, but not articulated and examined, is the idea that paths are created by human actors operating in time (Karnoe & Garud, 1998). Throughout this chapter, our evaluation of the applicability of path dependent arguments for explaining the dynamics of Silicon Valley integrates the creative dimension of path development.

Particularly important for understanding Silicon Valley is linking the opportunities technological evolution provided with the creation of institutions and even specific regional industrial cultures.² In effect, path dependency is intimately related with path creation. Silicon Valley's institutions and cultures can be understood as a set of evolving, path dependent routines nurturing specific combinations of extrafirm industrial patterns within its circumscribed region (Foray, 1991; Storper & Walker, 1988). Regional growth cannot be reduced to either business or technical developments; rather, technology and institutions dialectically create an unfolding path (Hirsch & Gillespie, chap. 3, this volume). Cook and Seely-Brown (chap. 1, this volume) use the "generative dance" as a metaphor for describing this dialectic.³

This chapter considers how the concept of *path dependency* can be extended to thinking about the creation and evolution of regions. The second section reviews the previous explanations for the growth and development of Silicon Valley. It also introduces our argument that to understand Silicon Valley, it is helpful to see it as two separate economies. The first economy consists of existing firms, whereas the second economy consists of the institutions that have evolved to nurture new startups. The third section examines the genesis of Silicon Valley as a high-technology region. The fourth section shows how it

² See Rao and Singh (chap. 9, this volume) for a discussion of the institutionalization of particular technologies. Especially interesting is their discussion of the biotechnology industry, which was a clear beneficiary of the Economy Two, which will be discussed.

³ *Ex poste facto*, the outcomes of this generative dance can be seen as a path or even a technological trajectory (on technological trajectories, see Dosi, 1984).

was the spin-off pattern developed among the semiconductor firms that catalyzed the industry, which was central to the creation of the institutions that now exist in Silicon Valley. The fifth section singles out venture capital as one critical institution for the development of Silicon Valley. The conclusion discusses the strengths and weaknesses of the path dependent perspective for explaining the creation and subsequent evolution of Silicon Valley.

PATH DEPENDENCY

The concept of path dependency was developed by economists such as Arthur (1988) and David (1986) to describe the phenomenon they noticed of apparently inferior technologies dominating market spaces (for a more detailed discussion, see Hirsch & Gillespie, chap. 3, this volume). Arthur (1994) developed abstract mathematical models showing how features such as increasing returns could create winner-take-all outcomes. David (1986, 1990, 1997, 1999), in a series of historical articles, demonstrated how this occurred in the adoption of specific technologies. They found that under certain conditions, early decisions reverberate through history, closing alternative paths and validating a single path. The implication is that history matters and outcomes need not be rational or optimal.

Though there have been a number of critiques of the claims of path dependency,⁴ this chapter considers path dependency as a significant contribution precisely because it problematizes the present. Path dependence accepts that small events can have very large later impacts. What is significant in this stance is that it permits these small events to be precipitated by noneconomic events. This is a critical opening for explanations not dependent on simple short-term profit maximizing. In a path dependent world, social constructions and strategic maneuvering in a nondeterministic environment are critical for path formation.⁵

It is not a great leap to accept that technology and the institutions in which it is embedded coevolve from path dependency (For the definitive discussion of evolutionary economics, see Nelson & Winter, 1982; on

⁴Critiques from the economic mainstream include Liebowitz and Margolis (1990, 1995). Sabel (1998) argued that in theoretical terms, path dependency removes human choice and collective action from the evolutionary process. One curious oversight in the QWERTY versus Dvorak debate is the fact that Dvorak was introduced in the 1930s. Dvorak had a much more difficult task, that is, dislodging an already established technology, QWERTY. A more comprehensive discussion and critique is provided by Ruttan (chap. 4, this volume) and Bassanini and Dosi (chap. 2, this volume).

⁵See Pinch (chap. 14, this volume) for an exposition of the social construction of technology perspective. We sympathize with the social constructionist perspective, but find the implicit deliberative and planned meaning of "construction" quite dangerous. Here, Cook and Seely-Brown's metaphor of a "generative dance" (chap. I, this volume) is closer to our perspective on how technologies and institutions evolve. There are no blueprints and no certainty.

embeddedness, see Granovetter, 1985). Hirsch and Gillespie (chap. 3, this volume) point out the nested, intertwined nature of many path dependent phenomena, that is, a particular technological choice is often an outcome of the interaction of a number of path dependent processes. Implicit, but not well developed, is the recognition that any path is, in fact, built by actors creating, using, and reshaping the infrastructure of institutions, routines, and organizations in which the technology is embedded. This activity includes suppliers, but ranges further to include capital equipment makers, specialized financial institutions, marketing and distribution organizations, educational institutions, and a myriad of other organizations, many of which are specialized in the needs of a particular industry. Often, perhaps more important, is creation on the demand side where occupying a market space and developing customers can be critical for the adoption of an innovation. Christensen (1992) showed in the hard disk drive (HDD) industry that the emergence of new customers was critical for the survival and growth of new entrants. At times, this also extends to creating new distribution, marketing, and retail networks. In other cases customers have to be mobilized; for example, what Lampe' (chap. 11, this volume) terms *technological spectacles* or *races* establish the characteristics of particular brands or even technological solutions (see also Rao & Singh, chap. 9, this volume).

Definitive explanations of industrial emergence and firm clustering in specific regions remain elusive. In a parallel to the path dependence and dominant design theories (Arthur, 1994; David, 1986; Henderson & Clark, 1990), industrial geographers such as Storper and Walker (1988) found that there are periods of locational opportunity before an industry clusters and locks-in into specific locations. Both economists dealing with innovations and industrial geographers studying regional industrial growth find that often there is an initial period of openness with a number of contenders prior to the selection of a dominant design or dominant location. It is at such moments that the small events can result in the long-term differences.

Arthur (1994) explicitly argued that positive feedbacks led to the clustering of the electronics industry in Silicon Valley. In an industry in which spin-offs are frequent, the industry will tend to cluster in a certain region when there are agglomeration economies. In the abstract model, the actual location selected is random and the spin-offs or agglomeration economies reinforce clustering in a particular region. This formal model is powerful, but even at the earliest stage, not all regions have equal competencies.

SILICON VALLEY-AN INTRODUCTION

Silicon Valley, as we shall see, had many antecedents and is the outcome of a pastiche of forces, accidents of history, planning, and human foresight. Al-

though integrated circuitry was the triggering technology for the development of the Silicon Valley agglomeration, it was not the only electronics technology in which it became a leader. For example, in the 1940s, Silicon Valley already had entrepreneurs experimenting with magnetic recording techniques at Ampex. But it was IBM's decision in 1952 to open a laboratory in San Jose to develop magnetic recording techniques for data storage that created the intellectual capital, which evolved into the merchant disk drive industry (Gan 1991). This magnetic recording expertise led not only to the HDD industry, but provided a basic technology for the telephone call processing equipment industry. Another important industry, computer networking, had its antecedents in the Xerox Palo Alto Research Center (PARC) laboratory and an Exxon-founded semiconductor firm, Zilog. Much of the knowledge that led to the relational database software industry was developed in the computer science department of IBM's San Jose Laboratory. What all these fields shared was extremely rapid technical change and quickly growing markets. The proliferation of new industry segments meant Silicon Valley could grow even faster than it would have had it been entirely dependent on the semiconductor industry.

For heuristic purposes, Silicon Valley can be conceptualized as two inter-related economies.¹ Economy One includes existing firms producing integrated circuits, software, computer networking equipment, computers, and a myriad of other electronics products, that is, the existing high-technology firms and other institutions, such as universities. *Economy Two* is a loosely structured network of venture capitalists, lawyers specializing in high-technology, accountants, and consultants. Their intention is to facilitate the creation and growth of firms that can later be sold to larger firms or listed on the stock exchange not to ship products. The introduction of this division is not so much driven by theoretical concerns, but rather for the optic it provides for understanding the dynamics of the region. The ability to develop new firms to exploit technological opportunities is dependent on different institutions than those necessary to operate already established firms.

The two economies are interrelated because Economy Two depends on Economy One. Conversely, firms successfully nurtured by the institutions of Economy Two become members of Economy One. However, these economies are not identical. In Economy One, the firms create products and services to be sold. In Economy Two, the product *is* firms, which embody a set of technologies and routines that another firm will purchase or that capital markets are willing to invest in by purchasing equity. In both cases, the pur-

¹This distinction is not theoretically driven. The purpose is to separate two quite different activities: the operation of existing firms and institutions and the operation of institutions dedicated to creating firms *de novo*. The point being that the dynamics of Silicon Valley are best understood through this analytic distinction, which is ungeneralizable to most other regions.

chases are justified by the belief that the firm will grow sufficiently to increase its value. In and of itself, innovatory activity in an already established firm does not directly benefit Economy Two of Silicon Valley. However, indirectly inventions and innovations in existing firms can be extraordinarily beneficial because they are the raw materials for new opportunities for entrepreneurship. Not surprisingly, Economy One firms are the single largest source of entrepreneurs for Economy Two.

When these two Economies are conflated or one is ignored, it is difficult to understand Silicon Valley. Regional growth in Silicon Valley is predicated on Economy Two. A suggestive study by Almeida and Kogut (1997; see also, Kogut, Walker, & Kim, 1991) using patent data and semiconductor firm location showed that the presence of the large semiconductor firms is highly correlated with the incidence of small firm establishments. They found that the large number of semiconductor firms (density) created the conditions for the establishment of still more firms. And, not surprisingly, the greatest density was in Silicon Valley. In effect, since the creation of Fairchild, the region developed a set of extrafirm institutions and routines that fueled growth and continuing reproduction. Of critical importance here is the intervening variable of an environment making large numbers of new entrants possible. This environment is composed of the institutions of Economy Two.

The dynamism of new firm creation and the wealth in Silicon Valley has drawn great interest from academics and policymakers. A number of explanations of the dynamics and operation of the Silicon Valley regional economy have been advanced. Saxenian (1994) explained Silicon Valley's success by comparing it to the relative stagnation of Route 128 in Boston during the 1980s. The heart of her argument is the proposition that Silicon Valley's firms remained flexible and interactive, whereas those established along Route 128 became hierarchical and rigid.⁷ She held that Route 128 firms were vertically integrated whereas Silicon Valley firms either remained, or wisely, decided to become specialists. There are difficulties with this comparison and explanation, as it ignores the fact that Route 128 was built on minicomputer systems, whereas Silicon Valley was built on the semiconductor component—a far more general or basic electronic part (Robertson, 1995). Silicon Valley cannot be reduced to existing firms or their interactions, rather any explanation of the dynamism must also explain the regional routines and institutions that nurture new firm formation.

There are also a number of cultural explanations of Silicon Valley (Weiss & Delbecq, 1990). Here, Economy Two is equated with a culture of entrepreneurship. Yet this provides little explanation of the economic and tech-

⁷I do not examine this argument here. For an analysis of this position, see Kenney (1998).

nological foundations for the culture and the concrete conditions that sustain it. In this particular formulation, *culture* refers to economic activity, that is, purposeful activity directed toward financial gain. Few of these cultural explanations connect individual's actions to their pursuit of financial gains. In effect, economic acts are explained by cultural attributes and potential economic explanations are downplayed (Weiss & Delbecq, 1990). As an example, the propensity to establish new firms is attributed to a culture of startups. Curiously, these explanations make no reference to the potential for the entrepreneur to secure large capital gains and the fact that the entire infrastructure of Economy Two has evolved to facilitate and share in those capital gains. The infrastructure does not encourage startups that have no potential for realizing large capital gains; in fact, venture capitalists even have a word for firms that do not go bankrupt, but cannot be sold—these are known as "zombies." This is a pejorative term for companies that are too small and have insufficient growth potential, but already have the venture capitalists' investments.

A less prominent stream of study has designated Economy Two as the critical feature of Silicon Valley. For example, Schoonhoven and Eisenhardt (1989) argued that Silicon Valley is an "incubator region," in which there are numerous institutions whose *raison d'être* is to nurture the establishment and growth of small startup firms aiming to exploit a market opportunity. They empirically tested the incubator region concept by examining the creation and survival of new semiconductor firms in the United States from 1978 to 1986. Their findings show that the Silicon Valley firms had greater survival rates. In other words, the social institutions existing in the region provided environmental resources that incubated these new firms. Florida and Kenney (1988a, 1988b) advanced the concept of a "social structure of innovation," by which we meant an interactive set of institutions dedicated to encouraging technological innovation.⁸ More recently, Bahrami and Evans (1995) conceptualized Silicon Valley as an "ecosystem" consisting of various institutions, skill sets embodied in individuals, and an entrepreneurial spirit. These three perspectives identify the ingredients that have made Silicon Valley so successful in creating new highly successful firms; however, they do not explicitly weave the trajectories of the technologies being exploited into their explanations.

The differences between firm organization in different regional industrial agglomerations can also be explained on another dimension, namely the technological dynamics the particular industry faces. For Robertson and Langlois (1995) the innovatory situation a region's industry faces in terms of

⁸Lynn, Reddy, & Aram (1996) advanced yet another somewhat similar concept of an "innovation community," however their concept is more general and seems to fit established industries better than it fits environments such as Silicon Valley or Route 128. Also, it is not quite as explicitly spatial.

the product cycle conditions the organization of the region's networks, interfirm interaction patterns, and firm structures. For example, the high-fashion garment district firms of Northern Italy face constant change in fashion designs, but these changes occur only along very limited dimensions, that is, in the designs, colors, fabrics, and shapes of the particular item; but the product, such as, jackets, pants, etcetera, does not change. In this situation, the change in production equipment and worker skills is gradual.

Silicon Valley faces a far more complicated set of changes including technologies, products, processes, and entire industrial categories. Not only are product generations rapid, but new product categories can emerge, even as other entire categories disappear (Kenney, 1998). This means turbulence is ongoing and interfirm and intrafirm relations are under continual stress. No firm or set of firms can be certain that its particular technology or product will survive. Requisite skills change rapidly, as established products evolve or are discontinued to be made in other regions (e.g., floppy disk drives and DRAMs), entirely new products are introduced constantly, customers and suppliers change, and new customers or suppliers emerge (for discussions of hard disk drives see Christensen, 1992; for RISC microprocessors, see Garud & Kumaraswamy, 1995; for LAN systems, see von Burg, 1998). New firms with superior technology can emerge rapidly and displace older firms committed to obsolete technologies.

The Langlois and Robertson (1995) thesis—that the industrial organization of regions and firms is correlated to the region's position on the product cycle—is an important contribution to understanding the linkage between particular types of networks and industrial regions. Their model explicitly recognizes the importance of the types of innovation or market changes facing firms, thereby incorporating technical change as a critical variable—a factor surprisingly underplayed in many explanations. For example, in the semiconductor industry, even while particular industrial segments such as DRAMs and microprocessors developed predictable trajectories and/or strong incumbent firms, new segments emerged, igniting a new cycle by lowering entry barriers and allowing new startups. Therefore, the semiconductor industry as a whole did not mature, rather the locus for new firm entry constantly shifted. In computer local area networking, there was a similar process of new firm formation at each discontinuity in the expansion of the network (von Burg, 1998). Economy Two is based on these technical and market discontinuities.

The economic dynamism of Silicon Valley is partly from the fast-growing established firms, such as Intel, Sun Microsystems, or Hewlett Packard, which have graduated to Economy One, but much more important are institutions of Economy Two that encourage new firm formation. Rapid firm formation is not unique to Silicon Valley; there have been periods and regions

before that experienced rapid firm formation. For example, Rao and Singh (chap. 9, this volume) discuss the early phases of the automobile industry. For autos, a dominant design emerged and new firm entry became quite difficult. The electronics (and also biotechnology) industries experienced repeated new opportunities, so even as certain segments developed a dominant design and entrenched firms, new segments opened up creating new spaces for new firm formation.

However, displacing explanation to a set of entities such as regional institutions does not really help us understand the development of Silicon Valley. Most observers treat economic institutions as natural phenomena that exist *sui generis*, when, in fact, they are created. Economic institutions and routines such as the venture capital investment process are the outcomes of complicated evolutionary paths, reinforced by success or diminished by failure. Obviously, the greater the success of such routines the more they were reinforced; in this way, they could eventually become an attribute of the "culture."

This section argued that there are two economies in Silicon Valley and both exhibit path dependent characteristics. The actual evolutionary process is quite complicated because both Economy One and Economy Two are moving along trajectories made possible by Moore's and Metcalfe's Laws, which postulate a world where value and capabilities are increasing so dramatically that new commercial opportunities are constantly being uncovered.⁹ The next section illustrates the path dependent nature of the semiconductor industry in Silicon Valley.

SEMICONDUCTORS

In 1947, the first operating semiconductor transistor was developed at Bell Laboratories in New Jersey.¹⁰ At that time, few foresaw the vast technological possibilities that the semiconductor's evolution would make possible. Semiconductors would permit the digitalization of many analog functions and quickly displayed an improvement curve that allowed a doubling of capacity approximately every 18 months driving the cost per transistor down dramatically. This meant that problems relating to insufficient or too expensive calculating capacity were constantly being solved, permitting a constant flow of new applications (i.e., watches, calculators, mobile phones, communications computers, ever smaller computers, and various

⁹Moore's Law states that the number of circuits that can be placed on a given area of silicon doubles roughly every 18 months (Moore, 1965). Metcalfe's Law states that for any number of n machines linked by a network you get n squared potential value (Gilder, 1993).

¹⁰For excellent discussions of the development of the semiconductor, see Braun and Macdonald (1982), Riordan and Hodgeson (1997).

other artifacts that contained embedded computing power). Analog functions and signals could be replaced by the increasingly sophisticated integrated circuitry, so records were replaced with compact disks, watch gears with a chip, typewriter gears and levers with word processing programs, or human hands and brains with computer controlled machine tools. The semiconductor eventually would allow physical phenomenon to be digitized that no one could have foreseen.

The pace of change confirmed by Moore's Law meant incessant change and the continuous emergence of new business opportunities to produce either inputs to integrated circuit production, integrated circuits optimized for various functions, or artifacts using integrated circuits. In this environment, obsolescence of artifacts, technologies, and capabilities was incessant, thereby opening space for new entrants (for a discussion of this, see Kenney, 1998; Kenney & Curry, 1998). In other words, market-dislocating technological advances occurred frequently. Yet, even more important, were innovations defining new markets and repeatedly semiconductors were critical enabling components for industries ranging from computer networking equipment to personal computers and mobile phones. For semiconductor companies, so many new business opportunities emerged that management had to decide which ones to pursue, recognizing that if a project was blocked internally, the engineers developing the technology might decide to use the knowledge to build a new firm (Intel, 1984).

Semiconductors were at the heart of a massive technological revolution (Gilder, 1989). The exponentially increasing processing power of integrated circuits permitted the creation of new products and the transformation of old products and industries. The per-unit price of information processing power embodied in integrated circuitry dropped at a 40% annual rate for more than 20 years. Braun and Macdonald (1982) provided the example of a Fairchild transistor sold in 1959 for \$19.75 that in 1962 was sold for \$1.80. Most significant, invariably the transistor was profitable at both prices as learning curve and mass production economies lowered costs and the design costs were amortized during the initial part of the learning curve. This incessant fall in prices meant that by 1997 the price per transistor on an integrated circuit chip dropped to less than \$.000001. These extraordinary price-learning curves created fantastic opportunities for increased profitability and constantly opened new business opportunities that could be exploited by startups (Gilder, 1989).

THE GENESIS OF SILICON VALLEY

Silicon Valley is a postwar phenomenon, however there were precursor firms in the prewar period. Sturgeon (2000) maintained that the Bay Area's pre-

World War II successes in developing vacuum tubes and a number of other devices formed the base on which the postwar growth was built. For example, in 1906, Lee de Forest invented the triode in the Bay Area and later in the 1920s, Philo Farnsworth relocated from Utah to the Bay Area with the intent of developing a working television. Farnsworth received financial support from W.W. Crocker, president of the Crocker National Bank (Fisher & Fisher, 1996). In the 1930s, Hewlett Packard was founded by two engineers at the behest of Frederick Terman, the dean of engineering and later provost at Stanford University (Leslie, 1993; Lowen, 1992). However, these disparate activities did not coalesce into a coherent pattern or set of practices.

World War II was a watershed for the U.S. electronics industry and the Bay Area benefitted greatly from massive electronics-related armaments spending. In Silicon Valley, a number of startups were established to take advantage of this spending. Also, defense contractors built a number of factories and research facilities in the area. During World War II, other smaller electronics firms were also established in the area (Sturgeon, 2000).

During World War II, Frederick Terman had gone to Boston to manage Radio Research Laboratories at Harvard University. After the war he returned more committed than ever to establishing an electronics industry in the Stanford vicinity. For the next 20 years, he encouraged major East Coast electronics firms to establish research and development (R&D) facilities close to Stanford University (Leslie, 1993). Also, he urged Stanford students, such as the Varian Brothers, to form electronics companies.

Terman's single most important intervention was convincing William Shockley, one of the coinventors of the transistor at Bell Laboratories, to return to his hometown, Palo Alto, and establish the startup firm, Shockley Semiconductor. There was an element of good fortune at play in luring Shockley back to Palo Alto. Shockley left Bell Laboratories and wanted to launch a firm to commercialize semiconductor technology. He approached a number of institutions on the East Coast, in particular, negotiating with Raytheon, an important transistor manufacturer, about funding his proposed startup. He demanded \$1 million and after a month of bargaining, Raytheon refused (Scott, 1974). He also negotiated with the Rockefeller venture capital division, but no agreement could be reached. After these failures, he began discussions with Arnold Beckman, the founder of Beckman Instruments in Los Angeles. They reached an agreement and Beckman funded Shockley to start a firm in Palo Alto (Riordan & Hoddeson, 1997).

Shockley's decision to locate in Palo Alto in itself was not significant, as Shockley Transistor never became an important firm. But, as fate would have it, a small significant event occurred. Shockley proved to be an ineffective manager, and eight of his engineers left to form Fairchild Semiconduc-