Evolutionary Theory and Endogenous Institutional Change

Danielle Allen

(James Bryant Conant University Professor, Harvard University),

Henry Farrell

(George Washington University, Political Science),

Cosma Rohilla Shalizi

(Carnegie Mellon Department of Statistics/Santa Fe Institute)

(alphabetical)*

Last LaTeX'd September 18, 2019

*We are grateful to Sam Bowles, Mark Blyth, Joshua Cohen, Federica Carugati, Ingrid Creppell, Avner Greif, Ellen Immergut, Nicolas Jabko, Dan Kelly, Robert Keohane, Nannerl Keohane, Jack Knight, Ruth Kricheli, Margaret Levi, Suresh Naidu, Josiah Ober, the late Elinor Ostrom, Melissa Schwartzberg, Adam Sheingate, Sven Steinmo, Kathleen Thelen, Maya Tudor and Barry Weingast for their comments on earlier versions of this paper, as well as workshop audiences at the European University Institute, the Max-Planck Institut fuer Gesellschaftsforschung, the Department of Political Science at Stanford University, Johns Hopkins University and the Center for Advanced Studies in the Behavioral and Social Sciences at Stanford University. At different points, Robert Keohane, Sven Steinmo and Kathleen Thelen gave particularly useful support. Henry Farrell gratefully acknowledges the support of the Woodrow Wilson International Center for Scholars. Cosma Shalizi acknowledges support from grants from the Institute for New Economic Thinking [[numbers]], the National Institutes of Health [[numbers]], and the National Science Foundation [[numbers]].

Contents

1 Introduction

Social scientists, including political scientists, economists, sociologists and anthropologists, have difficulty understanding the dynamic aspects of institutions. Their major theories stress stability, treating institutions as fixed game theoretic structures or equilibria (economics and political science), or as deeply shared common understandings (sociology and anthropology). Institutional change is usually treated as the product of exogenous forces, leading to serious theoretical questions about whether institutions have independent causal force, or are instead mere transmission belts between some external factor and actors' behavior [Przeworski, 2004]. Efforts to explain partly endogenous change are better at explaining how institutions can become self-reinforcing or self-undermining than at explaining change as such [Greif and Laitin, 2004].

This makes it harder for social science to address major puzzles. For example, social scientists broadly agree that the historical transition from closed and hierarchical societies to open societies is best understood as a process of dynamic institutional change [North et al., 2009a, Acemoglu and Robinson, 2013]. The difficulty of applying static understandings of institutions to dynamic processes has led scholars to casually invoke evolutionary concepts such as 'genetic drift' to explain long term processes of cumulative change [Acemoglu and Robinson, 2013], but they have not been able to move beyond broad analogy, or to understand how evolutionary forces might e.g. interact with power relations Levi [2013].

This article addresses this problem by providing a new and comprehensive framework for understanding institutional change as a dynamic evolutionary process. We treat institutions as populations of socially transmitted and roughly similar beliefs. Interaction between people with diverse viewpoints gives rise to evolutionary 'noise,' determining the rate at which new variant beliefs arise. Whether a belief becomes epidemic, leading to institutional change depends on network contagion, where variation in the topology of the network (the asymmetry of degree across nodes) captures the extent of power asymmetry in a given society, and the number of high degree nodes as a proxy for the extent to which elite actors favor or

disfavor a particular belief (as beliefs become more objectionable to powerful figures, high degree nodes are deleted). More generally, processes of network contagion can be treated as specific forms of 'large deviation' processes of change, allowing us to model institutional change as a stochastic-dynamical evolutionary process.

Our account has clear limitations. It is incapable of modeling 'mistakes' or more general processes of institutional backsliding and degeneration. Like other evolutionary accounts, it draws a sharp intellectual separation between the processes through which new variations are generated, and the processes through which they are diffused, while in actuality they are likely to be commingled. It does not provide a fully predictive account of institutional change, instead emphasizing that change will involve an irreducible element of stochasticity. Finally, it is better suited to explaining bottom-up processes of institutional change (in which variations well up from below and are selected through diffuse social processes), than top-down processes (in which leaders make the key decisions, and formal processes are crucial).

Even so, it provides a new set of tools that social scientists use to understand institutions. Most notably, it allows social scientists to compare dynamic processes of institutional change across different social settings. Societies will produce new possible institutional innovations (new beliefs) at different rates, depending on their viewpoint diversity (and the extent of interaction between those with different viewpoints). They will also vary in the rates at which new beliefs are taken up and become epidemic, and the *kinds* of beliefs that are taken up, depending on the extent of social power asymmetries.

To evaluate the plausibility of these claims, we apply these arguments to important recent social science controversies over institutional change in the Greek city-state system during the classical and Hellenic eras, comparing Athens and Sparta on the basis of qualitative data drawn from the historical record. Our approach captures the dynamic aspects of institutional change, providing a highly plausible explanation of why the two cities experienced such different broad trajectories of institutional change over the relevant period.

2 Institutions and Institutional Beliefs

Social scientists from different disciplines and approaches disagree over whether institutions have force because they simply provide information or shape deeper collective understandings. They broadly agree that institutions are rules that shape people's behavior because they instantiate individuals's beliefs about what other people believe, and how they may be expected to respond [Knight, 1992, Jepperson, 2002]. Social scientists furthermore usually presume that these beliefs either are unproblematically shared (so that everyone has the same beliefs about the institution). This implies that institutions are somehow transmitted unproblematically from individual to individual, a truly heroic assumption, given basic results from information theory, statistical learning theory, cognitive psychology and other disciplines.[1]

Our account, in contrast, models institutional change as a consequence of noisy transmission. Institutional beliefs - that is, beliefs about what other people believe is the appropriate rule in a given set of circumstances - may change as a product of communication and observation, so that new variant beliefs emerge. These variants may, or may not spread, again as a consequence of processes of social transmission.

This provides the basis of an *evolutionary* account of institutional change. There is much confusion within the social sciences about the implications of evolutionary approaches. Many social scientists are reluctant to acknowledge that evolutionary accounts as such can explain anything important about human social behavior. The publicity given to a noxious sub-literature in evolutionary psychology, whose dedication to crafting crude functionalist accounts of complex human phenomena is matched only by its incompetence at data analysis [Gelman, 2007], has done much to stroke this hostility. However, an evolutionary theory does not need to invoke biological adaptations in its explanations of particular phenomena, or make strong assumptions about some general guiding force [Padgett and Powell, 2012].

[1] Knight and North [1997] do note the problem, as do Greif and Laitin [2004] who assume, however, that individuals will rapidly converge on an excellent approximation of common knowledge.

All that an evolutionary account requires is that the theory of change invoke and describe the relationship between mechanisms of transmittable variation and selection.

Transmittable variation, in our account, involves the generation and transmission of new variant institutional beliefs, which differ markedly in some respect from the existing population of beliefs). People learn about institutions through observation and communication. New institutional beliefs will arise when individuals understand an institutional rule in a significantly different way to those around them. These beliefs may in turn be transmitted to others either through communication (when an individual explains her understanding of an institution to another), or through action (when a person behaves in ways that are at odds with others' understanding of a given institution, causing those others observing to question or re-evaluate their own understanding).

Thus, beliefs must be acquired through ordinary processes of learning, which are limited by the intrinsic limitations of any learning process, the biology of human memory, and the specific imperfections of particular forms of observation and communication. Such imperfections mean that individual beliefs are never perfectly transmitted, any more than other mental representations are [Sperber, 1996]. The background level of 'noise' in social transmission processes means that there is always some positive probability of new variant beliefs emerging.

However, the level of 'noise' may differ systematically across communities or social settings. A large body of findings across the social sciences suggests that the degree of *cognitive diversity* across individuals in a given social setting will have systematic consequences for the rate at which new variant beliefs will arise. By cognitive diversity, we mean the level of divergence across the schemas or broad systems of perspectives through which individuals understand the world around them. Schemas may be shaped by a variety of intrinsic and extrinsic factors that provide individuals with general toolkits through which to interpret their surroundings, including, for example, different religions, educational backgrounds, political philosophies and work experiences. These may in turn shape the particular beliefs

that individuals have about how a specific rule ought to be interpreted. The greater the variety of schemas in a given community or society, the greater the cognitive divergence.

Generally, we expect that the more contact there is between cognitively diverse individuals in a given community or society, the greater the likelihood of new institutional beliefs emerging in that community or society. Such contact will obviously lead to more disagreement about how an institution ought to be interpreted with respect to a given situation, leading to "mistakes" that may sometimes point in fruitful directions. It may also make it easier for individuals strategically and deliberately to contest and misinterpret institutions in ways that favor them [Hall and Thelen, 2009]. Greater cognitive diversity furthermore makes it more likely that individuals will be able to generate genuinely new insights by communicating among each other, learning from each others' point of view, and recombining the logics of different viewpoints to discover new possibilities [Page, 2007]. These three mechanisms, which roughly correspond to different major understandings of human communications across the social sciences (cultural, strategic and deliberative), differ sharply at the micro-level, but may reinforce each other in creating macro-level outcomes.

Institutional beliefs will be retained when they spread sufficiently widely to become epidemic across the population of actors, giving rise to a new institution. This can be understood as a standard process of *social contagion* across a social network comprising the individuals in a given community or society. Individuals will not be able to pass on a new institutional belief unless they are exposed to it, either through having others communicate the belief to them, or through observing others behaving in ways that accord with the new belief. They may then pass a more or less accurate version of this belief on to others with some positive probability, either through communicating it or behaving in accordance with it. This means that different network topologies will be associated with different probabilities of a social contagion spreading. Roughly speaking, the more connections there are, and the higher the variance there is in degree (so that some nodes in the network have many more connections than others), the more likely it is that a given contagion will spread. Power relations are

reflected by differences in 'degree,' or the number of links to a particular node, and have stark consequences for social contagion. Human societies tend to be characterized by social networks with high variance of degree, and also to involve assortative attachment between individuals of high degree [Newman and Park, 2003]. The higher the variance of degree, the more influential the highly linked nodes are in determining which contagions spread, and which die out.

This provides a straightforward way of understanding which institutional beliefs are likely to spread so that they become epidemic across the entire community, leading to institutional change, and which not. Variance in degree provides a rough index of the level of power asymmetry in a given society: the more variance there is, the more readily powerful actors (represented by nodes with high degree) can influence the process of institutional change. Institutional beliefs that are inimical to high degree nodes (they may not be in those actors' material interest, or they may seem inappropriate to those actors, or both) will be unlikely to be passed on by them, whether because they are blocked by implied or actual physical coercion, or by control over the legitimate channels of communication, or some combination. The more unequal power relations are in a given society, the less likely it is that institutional beliefs that are inimical to highly powerful actors will become endemic across the entire network. This is not to say that generally beneficial or attractive institutional beliefs cannot spread in societies with power asymmetries. However, they will typically only spread where they are not only broadly beneficial, but specifically congenial to powerful actors.[2]

Thus, an institution exists when the individual members of a community have institutional beliefs that are similar enough that they are roughly self-reproducing and mutually reinforcing over most situations most of the time. However, because social transmission processes are imperfect, and because people are likely (thanks to innate cognitive differences) to interpret institutions in somewhat different ways, institutional beliefs may change

[2] Our argument here is similar to Rodrik [2014, 197]'s observation that "Just as we think of technological ideas as those that relax resource constraints, we can think of political ideas as those ideas that relax political constraints, enabling those in power to make themselves (and possibly the rest of society) better off without undermining their political power."

in the process of transmission significantly enough so as to give rise to genuinely new beliefs. New beliefs have some positive chance of being taken up by others and transmitted (also imperfectly) across the relevant community. When a new and significantly different set of roughly congruent beliefs has spread widely enough to become endemic across the community, replacing the previous set, we may say that institutional change has occurred. We have moved from one set of roughly similar institutional beliefs to another. This process of change may be understood in evolutionary terms. New variant beliefs arise as a result of imperfect transmission, which is more likely to happen when there are high levels of contact between cognitively diverse individuals. These beliefs may be retained if they spread sufficiently widely to give rise to a new institution, but, depending on the level of power asymmetry, may be selected out if they are disfavored by powerful individuals.

3 Modeling Institutional Change

By decomposing institutions into populations of approximately shared beliefs across a given community of actors, and examining how variant beliefs are generated and selectively retained, we can begin to address Knight and North [1997, p. 214]'s challenge to model processes of institutional change that explain "how and why [human beings] develop theories in the face of pure uncertainty, [and] what makes those theories spread among a population and die out" Specifically, we can model institutional change as a specific application of a more general class of 'large deviations" or "stochastic-dynamical" approaches to understanding evolutionary processes. In stochastic-dynamical accounts, a noisy process draws specific new variations from an underlying possibility space. These variations may then spread across the population. Some variations may become firmly established in the population through "evolutionary drift," even though they do not confer any fitness advantages over existing traits. The common element of results across a very wide range of these models is that evolutionary processes obey *large deviations principles*, under which the probability of a certain

historical trajectory of strategies in the population is roughly exponential. These models of evolutionary processes identify three major factors — the rate at which new variations arise, the 'size' of the population across which a new variation has to spread, and the relative "difficulty" of potential trajectories to higher fitness states — as affecting the rate at which new evolutionary "fitness states" or "equilibria" emerge.

These models have not been typically applied to social or institutional evolution, because it has not been clear how well the mathematics travel. As we show in Appendices I and II, it is reasonable to characterize institutions as stable states, and the social contagion of institutional beliefs as a large deviations process, allowing us to apply this approach to understand how institutions evolve. First, we can capture how the "noise" processes described previously, as well as the "size" of the relevant population, affect the rate at which new variations are generated. We predict that societies with greater cognitive heterogeneity will have greater noise. Social structures may further affect the level of noise by affecting the likelihood that individuals with different cognitive approaches will come into contact with each other, thus changing the relevant "size" of the population that the belief has to spread across. Societies where individuals have heterogenous understandings of the world, but only rarely come into contact with each other, are obviously less likely to give rise to new beliefs than societies where cognitively diverse individuals are in frequent contact. Together, the level of heterogeneity and the extent of social contact among heterogenous individuals determine the speed with which evolutionary processes can explore the underlying possibility space by generating new variations. The higher the noise, the more quickly evolutionary processes can explore the relevant space.

Less abstractly, different environments will be more or less conducive to new institutional beliefs emerging. Social environments where there is plentiful institutional interaction and exchange between individuals with different cognitive maps of the world will have higher rates of variation (and hence, rates of possible institutional change) than social contexts where there is little such exchange. Contact between people with different skill sets and social

backgrounds may lead to the accidental or deliberate discovery of new institutional forms [Padgett and McLean, 2006]. Contact between diverse individuals will be more common in societies with internally heterogenous populations, engaged in different forms of economic and social activity with implications for their understanding of the world. It may also be produced by greater interchange with other societies, whether through temporary or permanent flows of people or trade.[3].

As already noted, power asymmetries, which can be represented as the variance of degree in a given social network mean that new institutional beliefs do not have an equal chance of being selected as the basis of new institutions and thus retained. Specifically, power asymmetries combined with different attitudes among powerful actors towards specific institutional beliefs may mean that beliefs that sit poorly with power elites are less likely to spread contagiously across the network. We capture this by assuming that if nodes of high degree belong disproportionately to a particular class which is inimical to a new belief, and so resist its spread, the new belief is being propagated over a network from which the high-degree nodes have been preferentially removed [Albert et al., 2000, Callaway et al., 2000]. Such "targeted" removal lowers the average degree of nodes on the effective network, as well as lowering the variance of degree; both of these make it harder for the innovation to spread, by raising the epidemic threshold. Preferentially blocking the high-degree nodes also changes the effective network for the diffusing innovation by increasing the average distance between nodes, especially between nodes which belong to highly transitive clusters. This is because the high-degree nodes are unusually apt to be long-range bridging ties (at least in many topologies, including the topologies most characteristic of human social ties), and in fact removing a fairly small fraction of the highest-degree nodes can effectively fragment the network into disconnected components. The implication of all this is when a new institutional

[3]See e.g. Mill [1909], "the economical advantages of commerce are surpassed in importance by those of its effects which are intellectual and moral. It is hardly possible to overrate the value, in the present low state of human improvement, of placing human beings in contact with persons dissimilar to themselves, and with modes of thought and action unlike those with which they are familiar. Commerce is now what war once was, the principal source of this contact."

belief is resisted by a group of actors of high degree, even without concerted collective action on their part, their individual resistance can block the belief from spreading. On the one hand, this can capture *variation in the resistance of powerful actors to specific institutional beliefs*. By taking a given topology and then preferentially removing high degree nodes from it, we can see how variation in the resistance of highly connected actors affects the likelihood that a new belief will become epidemic across the entire network leading to a new institution. On the other, we can compare the robustness of different network topologies (specifically, network topologies with higher or lower variance in degree) to node removal, and hence capture how the *interaction* between different levels of inequality of influence between elites and non-elites, and different levels of elite resistance shapes processes of institutional change.

Applying a stochastic-dynamical evolutionary approach to the contagion of institutional beliefs provides two major advantages over existing understandings of institutional change. First, it allows us to theorize the *rate at which new variations appear*. Specifically, it enables us to see how differences in the extent of cognitive diversity across the relevant community (and opportunities for people with different ways of thinking about the world to come into contact, converse with and observe each other) are likely to be associated with systematically different levels of institutional innovation. Societies where there is a high degree of contact among cognitively diverse people will produce variant institutional beliefs relatively frequently. Societies that are more homogenous, or that do not provide for rich contact between diverse individuals will be far less likely to produce such innovations.

Second, it allows us to theorize how different institutional beliefs will fare across different social settings, so that some have a relatively high probability of being selected, disseminating to produce new institutional stable states, while other beliefs have a very low probability. Specifically, it allows us to model how power disparities affect the spread of institutional beliefs. Beliefs that are uncongenial to socially influential actors are *ipso facto* less likely to spread across all plausible social networks. We can model the differences in action path faced by different beliefs by cumulatively deleting influential nodes. Uncongenial beliefs will

face an effective social network from which most influential nodes have been deleted (because the individuals occupying these nodes will not transmit the beliefs to others), and will be correspondingly unlikely to disseminate and generate new stable state institutions. Equally, however, different social settings may have greater or lesser power disparities (which we can represent by generating different network topologies). Topologies in which elite actors are less dominant (e.g. the differences between high degree nodes and ordinary nodes are less marked) will also be more resistant to node deletion, and hence less susceptible to the ability of elites to block uncongenial ideas. Taking these two together allows us to understand the rate of institutional change far better than before

4 Evidence from Athens and Sparta

As with other major approaches to institutional change [Greif and Laitin, 2004], our approach is difficult to test. It explicitly assumes that institutional change has an irreducibly stochastic element. Hence, we look instead to show how it can plausibly help resolve practical problems in the study of institutional change, applying our argument to the key cases of Athens and Sparta during the Classical and Hellenic eras.

Athens and Sparta have become paradigmatic historical cases for a highly influential current approach to institutional change and economic history [Carugati, 2019]. North et al. [2009b] argue that the most crucial distinction in economic history and today is between societies with 'limited access orders' which prevent non-elites from participating in organizational life and 'open access orders' which allow anyone to participate, leading Carugati et al. [2015], Ober and Weingast [2016] and others to characterize Athens and Sparta respectively as exemplary cases of open and limited access orders. However, this literature has difficulty in capturing the different institutional dynamics of open and limited access orders. On the one hand, treatments such as Ober and Weingast [2016], Carugati et al. [2015], characterize the institutions of both societies as stable equilibria in underlying games, where

institutions will last as long as the parameters (or quasi parameters [Greif and Laitin, 2004]) argue that external selective forces of inter-state competition guided their evolution. On the other, scholars like Ober [2009, 2008] invoke external selective pressures from the Athenian city-state system to argue that Athenian democracy flourished better than its competitors. The first set of approaches has difficulty in discussing dynamic processes of institutional change at all, while the second, by privileging external selective pressures, has difficulty in systematically describing the relationship between outside forces and the different internal institutional trajectories of the two cases, or of. This reflects the broader difficulty of institutional theorists across the social sciences in modeling the dynamics of institutional change.

Our account provides a systematic way of understanding the different trajectories over time of the two societies, plausibly demonstrating how differences in the cities' respective levels of cognitive heterogeneity and power asymmetries had consequences for their respective rate and kind of institutional change.

The two cases are roughly comparable. Ancient Athens and ancient Sparta were both "super-poleis" [Ober, 2008], subjected to much the same environmental pressures, and yet with different levels of heterogeneity and power asymmetry as described by contemporary observers.

There is strong evidence that Athens' level of cognitive heterogeneity - and contact between heterogenous individuals - was unusually high in the ancient world. Athens had a diverse internal economy, and its political structures actively encouraged contact between individuals from diverse backgrounds. Cleisthenes' reforms in 507 BCE created ten artificial tribes each of which was populated with one third coming from seaside villages, one third from urban neighborhoods, and one third from rural agricultural villages. The body that set the agenda for the assembly consisted of 50 men drawn by lot from each tribe. This council lived together for the year to guide policy in the city center; the chairmanship rotated from month to month from one tribe to the next. Both tribe and council brought

individuals from diverse backgrounds into sustained contact with each other. In addition, Athens engaged in extensive trade relations (the Athenian harbor in the Peiraeus was at the center of a lively Mediterranean economy) and had a substantial internal population of foreign permanent residents, or 'metics' as well as manumitted slaves. Foreign intellectuals attracted to the city included Herodotus, Protagoras, Gorgias, Lysias, and Aristotle, and at the annual performance of the tragedies, seats were reserved in the theater for foreigners. Although Socrates was notoriously executed for having introduced novel ideas about the gods to the city, this did not happen until he was 70. For most of the decades preceding his death, the Athenians tolerated his quite heterodox views, and continued after his death to tolerate contention between different schools of philosophy with competing perspectives.

In contrast, Sparta's internal cognitive heterogeneity and exposure to external perspectives was unusually low. Sparta sought both to reduce cognitive diversity among its citizens, and to control even more tightly such cognitive diversity as did arise. Sparta's famed education for its young citizens was homogenizing, even to the level of dress. They sang the same songs from Tyrtaeus continuously for several centuries. They were taught to spy on each other and report wrongdoing. They called themselves "homoioi" or "similar. The relatively very small population of citizens did indeed have regular contact with each other, but under commensal institutions engineered to promote a strong common culture among a group of individuals with a shared economic interest in maintaining an economy based on serf agriculture. After approximately 500 BCE (Cartledge 2002, 93), Sparta furthermore sought to discourage contact between citizens and foreigners for fear that alien ideas might spread and become established, deprecating trade and the use of money. After 500, Spartans begin to conduct ritual expulsions of non-elite foreigners (xenelasia) to avoid introducing corrupting ideas (Xen. RL 14A.4). There was a complementary proscription against travel abroad for Spartan citizens. Sparta's economy was primarily agrarian, relying on the work of a large body of helots, who were effectively slaves, and resident perioikoi, with limited rights. The ban on silver and gold and on citizens practicing trades, meant both that fewer foreigners

sought out Sparta than Athens, and that its citizenry, who were specialists in violence, had less occasion to interact with foreigners on account of sharing a craft or other business.

There is contemporary evidence that these differences in the level of cognitive heterogeneity had consequences for the spread of abilities and ideas. Thucydides, for example notes that while the Athenians were famously talkative—they lived in a world ringing with discourse — the Spartans were characteristically tight-lipped (the word 'laconic' comes from the archaic name for the region surrounding Sparta, Lakonia). In Pericles' famous funeral oration, Athens is described as "open to the world," while Sparta periodically deports foreigners. Records of Spartan debates provide relatively few examples of expression of alternative views.

There were similarly sharp differences in the level of power asymmetries across the two cities. With the exception of brief oligarchical interludes such as the rule of the Thirty, the Athenian political system provided equal rights to all male citizens while retaining subjugated positions for women and slaves. While elites dominated public speech (in general, known speakers were wealthy elite politicians, and older men were invited to speak first), they had to respond to the interests of a broader public of male citizens, allowing arguments and ideas to bubble up from a broader public. Solon's constitution instituted the rule of ho boulomenos, "whoever wishes," which in principle allowed anyone to speak at any time.

Sparta had far more limited citizenship, which depended both on descent and a stringent property requirement. It subjected the vast majority of Sparta's inhabitants to a reign of terror and surveillance intended to prevent revolt. Given that the helots and perioikoi were situated quite differently from the Spartan citizens, and given that the helots revolted three times, we can assume that they would have preferred quite different institutions from those that dominated. Sparta took aggressive and systematic action to control against selection of beliefs from these contexts. Spartan citizens conducted an annual war against the helots, and subjected them to campaigns of state-sponsored mass killings conducted by specialized troops or "Krypteia," the "hidden ones" (Ober 2015, 139). Helots were sometimes offered

the chance to become perioikoi - but might disappear if they accepted(Thucydides 4.80.3, cited in Ober 2105, 349n36).

These differences would lead us to expect that the two cities would be associated with very different trajectories of institutional change. First, there would be substantial differences in the rate at which *variations* are thrown up. New institutional beliefs would be generated relatively frequently in Athens, thanks to regular social intercourse both with the outside world (via trade and other relations) and among citizens coming from different backgrounds. In Sparta, in contrast, the rate of generation of new institutional beliefs would be substantially lower, thanks to far lower internal heterogeneity and contact with external viewpoints. Furthermore, there would be substantial differences in the kinds of variant institutional beliefs that were *selectively retained* to become epidemic, leading to institutional change. In the Athenian case, we would expect that institutional changes that were beneficial to a relatively diverse body of male citizens would have a relatively high chance of being selected (although these institutional changes would only accidentally benefit slaves, women or other excluded groups). In Sparta, in contrast, institutional beliefs would be extremely unlikely to be adopted unless they specifically benefited a small and elite group of citizen-landowners with very particular interests in the maintenance of a heavily asymmetric set of power relations.

Together, these mechanisms would predict that new institutional beliefs would arise more frequently in Athens, and that socially beneficial beliefs would have a high chance of disseminating, generating institutional change. In Sparta, in contrast, broadly beneficial institutional adaptations would be less likely to arise from ordinary social intercourse, and less likely to spread and lead to temporary institutional equilibria when they did arise.

These predictions accord with the observation of contemporary commentators such as Thucydides that rates of cultural and political change were much greater in Athens than in Sparta. Athens experienced seven major rounds of institutional reform, between 590 BCE and 322 BCE. The Spartans, in contrast, experienced major institutional change in the

period from 800 to 600 but then not again after that, until the collapse of their system of serfdom (helotage) in the wake of their 371 military defeat at Leuctra.

The timetable of major institutional change in Athens is as follows:

- 624 BCE: Draco's codification, and the organization of courts. 594/3 BCE: Solon ends debt slavery for Athenian citizens; introduces the assembly and makes prosecution and assembly speech open to anyone.

- 561-510 BCE: The Psistratid tyranny introduces thirty circuit judges and may have institutionalized tragedy and comedy as a civic ritual.

- 508/7 BCE: The reforms of Cleisthenes replace the four historic tribes with ten artificial tribes, introduces the Council of 500 as the agenda setting body for the assembly, and extends the court system.

- 462-460 BCE: Reforms led by Ephialtes and Pericles rein in the power of the elite court, introduce jury pay, open the chief magistracies to all but the poorest and limit citizenship to those with Athenian parents.

- 411/410 BCE: Oligarchical changes under the "rule of the 400," and then under the rule of the Five Thousand.

- 410/409 BCE: A commission is established to collect and publish all existing laws.

- 404/3 BCE: After Sparta defeats Athens in the Peloponnesian War, it installs the rule of the Thirty Tyrants in Athens.

- 403 BCE: Democratic resisters restore democracy, establishing amnesty, create a board of nomothetai to review the laws, establish a distinction between laws and decrees, introduce pay for service in the assembly and, by 380, constrain the power of generals to punish with death in the field.

- 338-332 BCE: The Treasury of the City acquires administrative pre-eminence; an effort is made to restore the powers of the Areopagus; new offices, like the "orderers" or kosmêtai and "moderators" or sôphronistai, are introduced, and military training for young men is intensified.

The timetable of major institutional change in Sparta is as follows:

- 776 BCE: Sparta and Elis found the Olympic Games.
- 735 BCE: Sparta conquers Messenia and establishes its system of serfdom (helotage), distributing land to all the Spartan citizens (numbering approx. 9000). 710 - 650 [?] BCE: Lycurgan reforms introduced the council of elders, which included the traditional two kings; allocated power to citizens to vote on proposals made by the Gerousia, introduced common education for Spartan children, a common mess for shared meals, the ban on silver and gold coinage, the ban on interactions with foreigners, the creation of the Crypteia, and, possibly, land redistribution and the ephors, chief magistrates who were on par to the kings.
- c. 495-90 BCE: Sparta introduces the practice of xenêlasia, or expulsions of foreigners. (Figueira 2003, p. 69).
- 409 BCE: King Pausanias tries to reduce the power of the ephors (Arist. Pol. 1301 b 17) and was expelled from Sparta.
- 400 BCE: A succession battle, which is won when Agesilaus accedes to the throne.
- 371 – 369 BCE: Sparta's citizen population is down to 1500 adult, male citizens, and Sparta is defeated by the Boeotians at Leuctra. This is followed by the liberation of the Messenian Helots, the foundation of Messene as an independent polis; and the collapse of the Spartan system of serfdom (helotage).

In short, a basic review of the timetable of institutional change confirms what any number of sources—Thucydides, Xenophon, Plato, Aristotle, and Plutarch—report. Institutional change in Athens was characterized by relatively frequent change (albeit not always for the better) and by frequent shifts towards a greater diffusion of benefits to male citizens (sometimes followed by efforts to roll them back). Spartan society, in contrast, was characterized by relative institutional stability, and the conservation of power and benefits to a small subset of the population.

In the Spartan case, there is a clear correlation between minimalized cognitive diversity, power structures that aggressively work against selection of alternative beliefs, and a relatively low rate of social and political change. Athens experienced greater political turbulence but also showed considerable resilience. It seems reasonable to assume that Athenian institutions tended to have benefits for a relatively broad public, as compared with its non-democratic competitors, although women remained without political voice, and their slave system was intact as of 322. Spartans, in contrast lived in a stable polity with no fundamental regime change for nearly three centuries; over the course of the last seventy years, however, of their status as a superpolis, their citizen numbers steadily declined. They were incapable of changing course to correct for that decline. By the time of the Battle of Leuctra, they no longer had enough citizens to populate their fighting force, at which stage they were forced to recruit perioikoi without the traditional Spartan training, undermining their traditional institutions. While both cities may be said to have 'flourished' in some very loose sense, Athens clearly produced more rapid innovations, that tended to benefit more of its citizens, albeit at the cost of substantially increasing instability.

These empirical findings are consonant with our framework's predictions, suggesting that an evolutionary approach can indeed help explain differences in the rate and kind of institutional change across different polities. The kind of long term stability that Carugati et al. [2015] argue are characteristic of Sparta are plausibly less a reflection of the nature of equilibria as such, than a specifically homogenous cultural environment, and enduring power

asymmetries that ensured that very few of the variations that were thrown up were likely to be selected. The experience of Athens, in contrast, demonstrates how high levels of heterogeneity can combine with less asymmetric power relations to produce more rapid changes that are, much of the time, likely to be beneficial to a broader subset of the population.

5 Conclusions

This article addresses a major besetting problem of institutional theory, showing how an evolutionary approach can model dynamic trajectories of institutional change, and capture variation across different social settings. Our approach provides a basic toolkit that can be applied in a variety of ways beyond the cases studied. Most obviously, it can be extended to understand other aspects of the historical transition from closed to open societies. Thus, for example, the close association between the diffusion of early Protestant arguments for reform in German speaking cities and the level of media competition [Dittmar and Seabold, 2015] can readily be understood in our framework, which may also plausibly be applied more generally to the diffusion of ideas in the Enlightenment [Mokyr, 2016]. So too, our approach may help clarify debates about the consequences of communications technology for autocracies and democracies [Farrell, 2012]. Specific applications within the social sciences include debates within political science over the time duration of institutional change [Hacker et al.], in cognitive anthropology over the diffusion and decay of transmitted cultural practices [Sterelny, 2017], in sociology over the relationship between networks and institutional diffusion [Clemens and Cook, 1999], and in economics over the persistence of apparently inefficient institutions over time [North, 1990]. While our approach will not resolve any of these debates it promises to make them more precise and tractable, by providing a general framework for modeling strategies, as well as disclosing other plausibly relevant causal factors that are not addressed systematically by existing approaches.

We also note the limitations of our approach. As our brief empirical account of Athens

and Sparta indicates, the openness of democratic systems to new changes may be a source of instability as well as adaptation - some innovations turn out to be harmful. Our modeling assumptions effectively preclude the possibility of broadly harmful institutions being adopted (while it is theoretically possible that evolutionary forces could lead to inferior institutions becoming established, it is wildly unlikely). However, by making this and other limiting assumptions explicit, while still providing tractable results, it serves as a spur to the development of better theories that could better address other urgent problems such as the instability of democracy.

A Mathematical Appendices

A.1 Large Deviations in Evolutionary Processes

Most readers will be familiar with the law of large numbers: if $Y_1, Y_2, \ldots Y_n$ are statistically-independent and identically distributed numerical variables, all with expected value μ, then as $n \to \infty$, $\frac{1}{n}\sum_{i=1}^{n} Y_i \equiv X_n \to \mu$. The most common sense of convergence here is "convergence in probability" or "stochastic convergence": you[4] can pick any margin of approximation η, and level of confidence δ, and I can find a (finite) number $N(\eta, \delta)$ so that $n \geq N(\eta, \delta)$ guarantees $\mathbb{P}\left(|\mu - X_n| > \eta\right) \leq \delta$. Turned around, given η and n, there is always a probability $\delta(\eta, n)$ of an η-sized deviation, which goes to zero as $n \to \infty$.

This usual form of the law of large numbers is true in great generality, but says nothing about the *rate* of convergence, i.e., the form of the function $\delta(\eta, n)$. The theory of large deviations[5] gives *quantitative* versions of laws of large numbers, focusing on the rate of

[4]We refer readers un-used to the idea of universal and existential quantifiers as moves in a two-person game to Hintikka [1996].

[5]This subject has spawned a large literature in probability theory and adjacent fields. The systematic modern definitions, terminology and notation is largely the work of Varadhan and collaborators (reviewed in Varadhan 1984, 2008). den Hollander [2000] is a fairly friendly introduction, presuming minimal knowledge of advanced probability. Dembo and Zeitouni [1998] is a deservedly-standard textbook which *does* presume measure-theoretic probability. Ellis [1985] explains the uses of the theory for the foundations of statistical mechanics in physics, as does, in less rigorous fashion, Touchette [2009]. For early applications to statistical inference, see Bahadur [1971].

convergence. Remarkably, such results also give (matching) lower bounds on the probabilities of such deviations from the expected behavior. Abstractly, such "large deviation principles" consider a family of random variables parameterized by a real number ϵ, say M_ϵ, with limit μ, and assert that

$$\mathbb{P}\left(M_\epsilon = x\right) \to \exp\left\{-\epsilon^{-1} J(x)\right\} \tag{1}$$

for some "rate function" $J(x) \geq 0$, which takes its minimal value of 0 at μ. A still more precise statement, which is what is usually used technically[6], is that, as $\epsilon \to 0$,

$$-\epsilon \log \mathbb{P}\left(M_\epsilon \in A\right) \to \inf_{x \in A} J(x) \tag{2}$$

The probability distribution thus concentrates around the point μ which minimizes the rate function, since the probability of visiting *any* set A not containing the minimizer μ shrinks exponentially — but fluctuations away from μ *do* happen, since those probabilities are *only* exponentially small, and not zero. The parameter ϵ thus gauges the extent of such fluctuations, or their importance compared to the limiting behavior.[7] Moreover, such a large deviation principle will apply conditionally: since, if $B \subset A$, $\mathbb{P}\left(M_\epsilon \in B | M_\epsilon \in A\right) = \frac{\mathbb{P}(M_\epsilon \in B)}{\mathbb{P}(M_\epsilon \in A)}$, one has $-\epsilon^{-1} \log \mathbb{P}\left(M_\epsilon \in B | M_\epsilon \in A\right) \to \inf_{x \in B} J(x) - \inf x \in A J(x)$ — "if something unlikely happens, it happens in the least unlikely way possible".

The exponential form of the probabilities that one obtains from the large deviation principle suggest the result of multiplying large numbers of independent events, and this is indeed one common route to an LDP. Roughly, if the outcome is a function of many statistically independent inputs, and each of these inputs has a smaller and smaller effect as ϵ shrinks, we may generically expect an LDP to hold.[8] (Obviously this is a *very* loose statement; the

[6] Actually, even this must be replaced by a slightly more complicated statement that handles some complications relating to the distinction between open and closed sets, and/or the possibility that the rate function J has dis-continuities. These details do not matter for our purposes.

[7] In the case of sample means and the basic law of large numbers, one usually takes $\epsilon = 1/n$. In fact, one usually obtains a rate function of the form $J(x) \propto (x - \mu)^2$, at least for x near μ, as might be anticipated from the central limit theorem.

[8] The usual counter-example, where a large deviation principle does *not* hold, is to consider a sequence where Y_1 is picked randomly from an arbitrary distribution, with mean μ, but thereafter all $Y_n = Y_{n-1}$. The

cost of precision is a lot of technicality that is un-illuminating for our present subject.) One area where these conditions apply is in stochastic models of learning in games. Here the point is not that different players are statistically independent — they are not! — but that each player, at each move, is at least *partially* subject to independent random influences. So long as no one player, or move, has too much of an effect over the course of the game, we can anticipate that an LDP will hold, and such results have been rigorously established for large cases of models.

One advantage of this rather abstract formulation, however, is that the same machinery can work in situations very different from that of simple sample averages. For example, the M_ϵ might be whole trajectories of stochastic processes, i.e., each realization x would be a *function* $[0,T] \mapsto \mathbb{R}^d$. We can then ask whether these stochastic processes converge on to some limiting process as $\epsilon \to 0$, and, when an LDP applies, we would know that (i) the most likely trajectory is the one which minimizes the rate function, i.e., the x^* such that $J(x^*) = 0$, and (ii) trajectories x which deviate from x^* become exponentially more and more unlikely, as $J(x)$ grows. In these cases, the rate function $J(x)$ typically takes the form of what is called, by analogy to physics, an "action", i.e., an integral

$$J(x) = \int_{t=0}^{T} L(x(t), \dot{x}(t))dt \tag{3}$$

where L is some function of both the position of the trajectory and its velocity or first derivative (or possibly higher derivatives, etc.), so that the convergence to the most-likely trajectory x^* is also a "principle of least action"[9]. When this is the case, we know from the calculus of variations [Boas, 1983, §9.5] that the optimal trajectory x^* will solve the ("Euler-Lagrange") differential equations

$$\frac{\partial L}{\partial x_i} - \frac{d}{dt}\frac{\partial L}{\partial \dot{x}_i} = 0 \tag{4}$$

sample mean does not converge to μ, because the initial random draw never becomes any less influential.

[9]Whether something like a large deviations principle is the origin of the principle of least action in physics is an interesting question, but gets into technical issues which aren't relevant here at all [Eyink, 1996].

which resolves into a system of ordinary differential equations for the evolution of $x^*(t)$.

To make all this more concrete, we will consider a series of increasingly elaborate examples (§§A.1.1–A.1.2), building towards a general result (§A.1.3) about the behavior of "population games", in which agents of multiple types receive pay-offs which are functions of the distribution of actions over the population, and can change their actions in response to those pay-offs. (The details will be made precise below.) This can be handled, at least non-rigorously, by mathematically elementary tools, and doing so illustrates the kinds of considerations that go into establishing LDPs for more complicated processes. This lets us treat the problem of transitions between stable states of population processes in considerable generality (§A.1.6). We build up to this, however, by first considering LDPs for independent, binary events, and for a simple model of the transmission of diseases or social information.

A.1.1 LDP for Binary Sequences

We start by considering a sequence of binary outcomes, so each Y_i is either 0 or 1, and $X_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the proportion of 1s in the first n variables. We assume that these binary variables are statistically independent, with the common probability $\mathbb{P}(Y_i = 1) = p$. What is the probability that $X_n = x$, for any particular x? By assumption, nX_n has a binomial distribution, so

$$\begin{aligned}
\mathbb{P}(X_n = x) &= \mathbb{P}(nX_n = nx) && (5)\\
&= \binom{n}{nx} p^{nx}(1-p)^{n-nx} && (6)\\
&= \frac{n!}{(nx)!(n-nx)!} p^{nx}(1-p)^{n-nx} && (7)
\end{aligned}$$

We're interested in the exponential decay rates of probabilities, so we take logarithms and divide through by n:

$$\frac{1}{n}\log \mathbb{P}(X_n = x) = \frac{1}{n}\left(\log n! - \log (nx)! - \log (n-nx)!\right) + x\log p + (1-x)\log(1-p) \quad (8)$$

Since we're interested in what happens at large n, we use Stirling's approximation, that $\log n! \approx n \log n$:

$$\begin{aligned}
\frac{1}{n} \log \mathbb{P}\left(X_n = x\right) &\approx \log n - x \log nx - (1-x) \log (n - nx) + x \log p + (1-x) \log (1-p) && (9)\\
&= -x \log x - (1-x) \log (1-x) + x \log p + (1-x) \log (1-p) && (10)
\end{aligned}$$

The first two terms, which are independent of the true probability p, are what's left of the combinatorial factor $\binom{n}{k}$; they indicate that some proportions (those near $x = 0.5$) can be realized in a very large number of ways, and occupy a large region in the space of possible outcomes[10]. The other two terms, involving both x and p, are about the mis-match between what the system "wants" to do (p) and the outcome we're considering (x). We can combine terms to get

$$\frac{1}{n} \log \mathbb{P}\left(X_n = x\right) \approx -x \log \frac{x}{p} - (1-x) \log \frac{1-x}{1-p} - \equiv J(x,p) \quad (11)$$

which one can check is maximized, at 0, by taking $x = p$. (The quantity on the right-hand side is the **(Kullback-Leibler) divergence** between a binary distribution with probability x and one with probability p, also called the "relative entropy", or "expected normalized log-likelihood".) Since n has disappeared from the right-hand side, we also have the limit:

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}\left(X_n = x\right) = -J(x,p) \quad (12)$$

Finally, we can get the probability that X_n falls into some range of outcomes B: since

$$\mathbb{P}\left(X_n \in B\right) = \int_{x \in B} \mathbb{P}\left(X_n = x\right)$$

[10] In fact, $-x \log x - (1-x) \log (1-x)$ is the **entropy** of a binary distribution with probability x, so high-entropy outcomes are favored, all else being equal.

and each term in the integral is decaying exponentially as n grows, the integral will also shrink exponentially, but with the slowest achievable decay rate[11]:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(X_n \in B\right) = -\min_{x\in B} J(x,p)$$

Obviously, this is not a rigorous demonstration[12], but the deficiencies can be patched up, at some cost in technicalities.

The *kind* of result also generalizes: rate functions generally combine two terms, one of them purely combinatorial, gauging the size of the region in state space compatible with a given outcome, and another, which indicates how much points in that region are favored or disfavored by the dynamics themselves. The sum of these two parts, the over-all rate-function, is generally a divergence between two distributions.

LDP for Multinomial Sequences Essentially the same derivation works if the Y_i are not binary, but take values in any finite space, independently and with a common distribution p. If we make X_n the sample distribution of $Y_1, \ldots Y_n$ (also called the "empirical distribution"), then

$$\mathbb{P}\left(X_n = x\right) = (n!) \prod_{j=1}^{k} \frac{p_j^{nx_j}}{(nx_j)!} \tag{13}$$

where k is the number of categories or possible values, $p_1, \ldots p_k$ are their true probabilities, and $x_1, \ldots x_k$ are the sample proportions. The math then follows along exactly the same lines as for the binary, $k = 2$ case. The final result, sometimes called "Sanov's theorem", is

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(X_n \in B\right) = -\min_{x\in B} J(x,p) \tag{14}$$

[11] This is "Laplace's principle" for approximating exponential sums or integrals.

[12] What if nx isn't an integer? How do we know a sum of exponentially-decaying quantities is also exponentially decaying? Etc., etc.

where now the divergence is

$$J(x,p) = \sum_{j=1}^{k} x_j \log \frac{x_j}{p_j} \tag{15}$$

A.1.2 LDP for the SI Epidemic Model

We next consider a very basic model of contagion or social influence, where a population of n agents is divided into two types (or "compartments"), S for "susceptible" and I for "infectious"[13]. The one transition possible is that a susceptible individual can acquire the infection from an infectious individual. (As a reaction, this would be written $S + I \to 2I$.) Time proceeds in discrete steps of length h, and the infectiousness rate is ρ, so the probability of infection of a generic S is $\rho(I(t)/n)h$. Define $X(t)$ to be $(S(t)/n, I(t)/n)$, so the infection probability is $\rho(x(t))h$. One typically assumes that the more individuals are infected, the greater the probability of encountering one, so that $\rho(x(t)) = rx_2(t)$ for some constant r. This assumption is not, however, essential to what follows.

Conditional on $X(t) = x(t)$, the distribution for the number of new infections is binomial, with $n - I(t) = nx_1(t)$ trials, and success probability $h\rho(x(t))$. (We return to the assumption that the infection rate is proportional to the *density* of the infected in §A.1.5 below.) The number of new infections, in turn, determines both $S(t+h)$ and $I(t+h)$.

We can now work out the transition probabilities, much as we worked out probabilities

[13] In the jargon, this is an "SI" model. More elaborate models add more types or compartments, e.g., R for "recovered" (or "removed"), leading to SIR models, or allow for more transitions, e.g., SIS models in which infected agents can revert to being susceptible. Models of this form are extensively used in epidemiology and ecology, and have been ever since pioneering work in the early 20th century. (It is conventional these days to attribute this to the work of Sir Ronald Ross on malaria, but early reviews (e.g., Lotka 1924, ch. VIII, pp. 79–82) make it clear that this was much more of a collective effort.) For a textbook treatment, see Ellner and Guckenheimer [2006].

for binary sequences above:

$$\mathbb{P}\left(X(t+h)=x(t+h)|X(t)=x(t)\right) \tag{16}$$
$$= \binom{nx_1(t)}{n(x_2(t+h)-x_2(t))}(h\rho(x(t)))^{n(x_2(t+h)-x_2(t))}(1-h\rho(x(t)))^{n(x_1(t)-(x_2(t+h)-x_2(t)))}$$

$$\log\mathbb{P}\left(X(t+h)=x(t+h)|X(t)=x(t)\right) \tag{17}$$
$$= \log\left(nx_1(t)\right)!-\log\left(n(x_2(t+h)-x_2(t))\right)!-\log\left(n(x_1(t)-(x_2(t+h)-x_2(t)))\right)!$$
$$+n(x_2(t+h)-x_2(t))\log h\rho(x(t))+n(x_1(t)-(x_2(t+h)-x_2(t)))\log\left(1-h\rho(x(t))\right)$$
$$\approx nx_1(t)\log n+nx_1(t)\log x_1(t) \tag{18}$$
$$-n(x_2(t+h)-x_2(t))\log n-n(x_2(t+h)-x_2(t))\log\left(x_2(t+h)-x_2(t)\right)$$
$$-n(x_1(t)-(x_2(t+h)-x_2(t)))\log n$$
$$-n(x_1(t)-(x_2(t+h)-x_2(t)))\log\left(x_1(t)-(x_2(t+h)-x_2(t))\right)$$
$$+n(x_2(t+h)-x_2(t))\log h\rho(x(t))$$
$$+n(x_1(t)-(x_2(t+h)-x_2(t)))\log\left(1-h\rho(x(t))\right)$$
$$= n(x_1(t)-(x_2(t+h)-x_2(t))-(x_1(t)-(x_2(t+h)-x_2(t))))\log n \tag{19}$$
$$+nx_1(t)\log x_1(t)$$
$$+n(x_2(t+h)-x_2(t))\log\frac{h\rho(x(t))}{x_2(t+h)-x_2(t)}$$
$$+n(x_1(t)-(x_2(t+h)-x_2(t)))\log\frac{1-h\rho(x(t))}{x_1(t)-(x_2(t+h)-x_2(t))}$$

$$\frac{1}{n}\log\mathbb{P}\left(X(t+h)=x(t+h)|X(t)=x(t)\right) \tag{20}$$
$$\approx x_1(t)\log x_1(t)$$
$$+(x_2(t+h)-x_2(t))\log\frac{h\rho(x(t))}{x_2(t+h)-x_2(t)}$$
$$+(x_1(t)-(x_2(t+h)-x_2(t)))\log\frac{1-h\rho(x(t))}{x_1(t)-(x_2(t+h)-x_2(t))}$$

Combine terms with common non-log pre-factors, so

$$\begin{aligned}
&\frac{1}{n}\log\mathbb{P}\left(X(t+h)=x(t+h)|X(t)=x(t)\right) \qquad (21)\\
&\quad\approx\; x_1(t)\left[\log x_1(t)+\log\left(1-h\rho(x(t))\right)-\log\left(x_1(t)-(x_2(t+h)-x_2(t))\right)\right]\\
&\qquad +(x_2(t+h)-x_2(t))\left[\log\frac{h\rho(x(t))}{x_2(t+h)-x_2(t)}-\log\frac{1-h\rho(x(t))}{x_1(t)-(x_2(t+h)-x_2(t))}\right]
\end{aligned}$$

As we let $h\to 0$, we have $x_2(t+h)-x_2(t)\to h\dot{x}_2(t)=-h\dot{x}_1(t)$. (We can restrict ourselves to continuous trajectories because we are also letting n grow, so we can't have very large jumps, i.e., discontinuities.) Furthermore, using the fact that $\log(1+a)\approx a$ for small a, we get

$$\begin{aligned}
&\frac{1}{n}\log\mathbb{P}\left(X(t+h)=x(t+h)|X(t)=x(t)\right) \qquad (22)\\
&\quad\approx\; x_1(t)\left[\log x_1(t)-h\rho(x(t))-\log\left(x_1(t)\left(1-\frac{h\dot{x}_2(t)}{x_1(t)}\right)\right)\right]\\
&\qquad +h\dot{x}_2(t)\left[\log\frac{\rho(x(t))}{\dot{x}_2(t)}+h\rho(x(t))+\log\left(x_1(t)\left(1-\frac{h\dot{x}_2(t)}{x_1(t)}\right)\right)\right]\\
&\quad=\; x_1(t)\left[-h\rho(x(t))-\log\left(1-\frac{h\dot{x}_2(t)}{x_1(t)}\right)\right] \qquad (23)\\
&\qquad +h\dot{x}_2(t)\left[\log\frac{\rho(x(t))}{\dot{x}_2(t)}+\log x_1(t)+h\rho(x(t))+\log\left(1-\frac{h\dot{x}_2(t)}{x_1(t)}\right)\right]\\
&\quad=\; hx_1(t)\left[-\rho(x(t))+\frac{\dot{x}_2(t)}{x_1(t)}\right] \qquad (24)\\
&\qquad +h\dot{x}_2(t)\left[\log\frac{\rho(x(t))x_1(t)}{\dot{x}_2(t)}+h\rho(x(t))-h\frac{\dot{x}_2(t)}{x_1(t)}\right]
\end{aligned}$$

Since we are letting $h\to 0$, second-order terms in h become negligible, so we can drop the last two terms on the right-hand side, getting

$$\begin{aligned}
&\frac{1}{n}\log\mathbb{P}\left(X(t+h)=x(t+h)|X(t)=x(t)\right) \qquad (25)\\
&\quad\approx\; h\left(\dot{x}_2(t)-x_1(t)\rho(x(t))+\dot{x}_2(t)\left[\log\frac{\rho(x(t))x_1(t)}{\dot{x}_2(t)}\right]\right)
\end{aligned}$$

Since transitions at each time step are independent of previous transitions given the

current state, if we want the probability of a whole trajectory $x : [0, T] \mapsto \mathbb{R}$, we would multiply the probabilities, i.e., add the log probabilities. Hence

$$\begin{aligned}
&\frac{1}{n} \log \mathbb{P}\left(X = x\right) && (26)\\
&\approx \sum_{i=0}^{T/h} h \left(\dot{x}_2(ih) - x_1(ih)\rho(x(ih)) + \dot{x}_2(ih) \log \frac{\rho(x(ih))x_1(ih)}{\dot{x}_2(ih)} \right) \\
&\rightarrow \int_{t=0}^{T} dt \left(\dot{x}_2(t) - x_1(t)\rho(x(t)) + \dot{x}_2(t) \log \frac{\rho(x(t))x_1(t)}{\dot{x}_2(t)} \right) && (27)
\end{aligned}$$

Since, by convention, the rate function is the *negative* of the rate of exponential probability decay, we have

$$J(x) = \int_{t=0}^{T} dt \left(\rho(x(t))x_1(t) - \dot{x}_2(t) + \dot{x}_2(t) \log \frac{\dot{x}_2(t)}{\rho(x(t))x_1(t)} \right) \quad (28)$$

Thus, the rate function is an integral along the trajectory, as promised, with the integrand being a function of both $x(t)$ and $\dot{x}(t)$, i.e.,

$$\begin{aligned}
J(x) &= \int_{t=0}^{T} L(x_2(t), \dot{x}_2(t))dt && (29)\\
L(u, v) &= \rho(u)(1-u) - v + v \log \frac{v}{\rho(u)(1-u)} && (30)
\end{aligned}$$

(Remember that $x_1(t) = 1 - x_2(t)$.) It is straight-forward to check[14] that $L(u, v) = 0$ when and only when $v = \rho(u)(1-u)$, and moreover that $L(u, v) \geq 0$. Thus, the action-minimizing trajectory is the trajectory x^* which solves the ordinary differential equation

$$\dot{x}_2(t) = \rho(x(t))(1 - x_2(t)) \quad (31)$$

But this is well-known to be the ODE version of the SI model [Ellner and Guckenheimer,

[14]The first assertion follows by direct calculation, which also shows that this is a local minimum (in v). For the second, observe that since $v \log v$ is convex, $L(u, v)$ is convex in v for each u, hence the local minimum at $v = \rho(u)(1-u)$ is also a global minimum (for each u).

2006, chapter 6].

A.1.3 LDP for General Population Processes

The SI model we've just considered is a member of a much broader family of what are sometimes called **population processes**. In these, a population of agents is divided into a finite number k of distinct types or classes, but agents can switch types, with probabilities that are a function of the current over-all distribution of the population across types. This very broad class of models includes a great many commonly-used models of social learning and change. Much of evolutionary game theory, for instance, concerns "population games" [Sandholm, 2010], where each agent chooses a strategy (= type), receives a pay-off which depends on the distribution of strategies, and changes strategies in response to its payoff. A slight extension makes changes of strategy a function of the recent history of the game, and have been used to model the evolution of institutions and the self-organization of social conventions [Young, 1998].

In this section, we will work out an LDP for quite generic population processes. The argument will work very similarly to that for the SI model, but with one important additional complication. In the SI model, we could identify the rate of change in the proportion of infectious agents, $\ldots X_2(t)$, with the rate of transitions from susceptible to infectious, because that was the only transition. In general, however, the *net* rate of change for type j, $\dot{X}_j(t)$, will be a sum of the *gross* transition rates to type j from all other types, minus the gross transitions from type j to all other types. It will turn out that we will need to track all these transitions, at least as an intermediate step.

Here then is the general case. There are k types of agent, so $X(t)$ (or its realization, $x(t)$) is a k-dimensional vector, though confined to the simplex, i.e., the region where $x_i(t) \geq 0$, $\sum_i x_i(t) = 1$. Over a time-step of length h, the probability that any one agent of type i changes to type $j \neq i$ is $h\rho_{ij}(x)$. Introduce the variables Y_{ij} to be the *gross* fluxes from type i to type j, i.e., changes in proportion per unit time. Clearly, if we know all the Y_{ij}, we

can determine $X(t+h)$, since $X_i(t+h) = X_i(t) - h\sum_{j\neq i} Y_{ij} + h\sum_{j\neq i} Y_{ji}$. For convenience, define $y_{i\cdot} \equiv \sum_{j\neq i} y_{ij}$, and similarly $\rho_{i\cdot} \equiv \sum_{j\neq i} \rho_{ij}$.

Now, let's look at the large deviations of Y given X.

$$\frac{1}{n}\log \mathbb{P}\left(Y=y|X=x\right) \tag{32}$$

$$= \frac{1}{n}\log \prod_{i=1}^{k} \frac{(nx_i)!}{\left(\prod_{j\neq i}(nhy_{ij})!\right)(nx_i - nhy_{i\cdot})!}\left(\prod_{j\neq i}\left(h\rho_{ij}(x)\right)^{nhy_{ij}}\right)\left(1-h\rho_{i\cdot}(x)\right)^{nx_i - nhy_{i\cdot}}$$

$$\approx \frac{1}{n}\sum_{i=1}^{k} nx_i \log nx_i - \sum_{j\neq i} nhy_{ij}\log nhy_{ij} - (nx_i - nhy_{i\cdot})\log(nx_i - nhy_{i\cdot}) \tag{33}$$
$$+nh\sum_{j\neq i} y_{ij}\log h\rho_{ij} + n(x_i - hy_{i\cdot})\log(1-h\rho_{i\cdot})$$

$$= \sum_{i=1}^{k}(\log n)\left(x_i - h\sum_{j\neq i} y_{ij} - (x_i - hy_{i\cdot})\right) + x_i\log x_i + h\sum_{j\neq i} y_{ij}\log\frac{h\rho_{ij}(x)}{hy_{ij}} \tag{34}$$
$$-x_i\log(x_i - hy_{i\cdot}) + hy_{i\cdot}\log(x_i - hy_{i\cdot}) + x_i\log(1-h\rho_{i\cdot}(x)) - hy_{i\cdot}\log(1-h\rho_{i\cdot}(x))$$

$$\xrightarrow{h\to 0} \sum_{i=1}^{k} x_i\log x_i - x_i\log x_i + x_i\frac{hy_{i\cdot}}{x_i} - hx_i\rho_{i\cdot} \tag{35}$$
$$+hy_{i\cdot}\log x_i - \frac{h^2 y_{i\cdot}^2}{x_i} + h^2 y_{i\cdot}\rho_{i\cdot}(x) + h\sum_{j\neq i} y_{ij}\log\frac{\rho_{ij}(x)}{y_{ij}}$$

$$\to h\sum_{i=1}^{k} y_{i\cdot} - x_i\rho_{i\cdot}(x) + (\log x_i)\sum_{j\neq i} y_{ij} + \sum_{j\neq i} y_{ij}\log\frac{\rho_{ij}(x)}{y_{ij}} \tag{36}$$

$$= h\sum_{i=1}^{k}\sum_{j\neq i}(y_{ij} - \rho_{ij}(x)x_i) + y_{ij}\log\frac{\rho_{ij}(x)x_i}{y_{ij}} \tag{37}$$

using the tricks that, for small h, $\log(1-h\rho_{i\cdot}(x)) \approx -h\rho_{i\cdot}(x)$; that, for small h, $\log(x_i - hy_{i\cdot}) \approx \log x_i - hy_{i\cdot}/x_i$; that terms of order h^2 are comparatively negligible; and that (regardless of h) we can re-expand $y_{i\cdot}$ and $\rho_{i\cdot}$ as sums of y_{ij} and ρ_{ij}.

Accordingly, the rate function for a continuous trajectory (x, y) is

$$J(x,y) = \int_0^T dt \sum_{i=1}^{k} \sum_{j \neq i} (\rho_{ij}(x(t))x_i(t) - y_{ij}(t)) + y_{ij}(t) \log \frac{y_{ij}(t)}{\rho_{ij}(x(t))x_i(t)} \quad (38)$$

$$= \int_{t=0}^T dt \sum_{i=1}^{k} \sum_{j \neq i} L_{ij}(x(t), y_{ij}(t)) \quad (39)$$

$$L_{ij}(u,v) \equiv \rho_{ij}(u)u_i - v + v \log \frac{v}{\rho_{ij}(u)u_i} \quad (40)$$

where it is again easy to check that each $L_{ij} \geq 0$, and that the minimum of 0 is attained uniquely when $v = \rho_{ij}(u)u_i$. (It should be understood that the rate-function is ∞ if the fluxes do not net out, i.e., $\sum_i \sum_{j \neq i} y_{ji}(t) - y_{ij}(t) \neq 0$, or, if any of the fluxes are negative.) The rate-minimizing trajectory, then, has both

$$y^*_{ij}(t) = \rho_{ij}(x^*(t))x^*_i(t) \text{ and} \quad (41)$$

$$\dot{x}^*_i(t) = \sum_j y^*_{ji}(t) - y^*_{ij}(t) \quad (42)$$

(Note that this is what we would get, after a lot of manipulation, if we invoked the Euler-Lagrange equation, Eq. 4.)

As discussed above, in the warm-up case of the SI model, the number of gross transitions from S to I, which is what Y_{12} would track, is also equal to the *net* transitions, which is what $\dot{X}_2$ tracks (or $-\dot{X}_1$), so we did not need the Y variable. But in general, when there are multiple ways into or out of each type, to get a simple form of the rate function, we need the auxiliary, flux variables Y tracking gross transition rates.

This may seem unfortunate, since we usually don't care about the augmented process, with the gross fluxes, but just the X process. However, here we can appeal to a general result of large deviations theory called "the contraction principle" [den Hollander, 2000]: if

Z obeys the LDP with rate function $J(z)$, then $U = f(Z)$ obeys the LDP with rate function

$$\inf_{z:f(z)=u} J(z)$$

Accordingly, the probability of seeing a trajectory x will be

$$\begin{aligned} & -\frac{1}{n}\log \mathbb{P}\left(X=x\right) \\ & = \inf_{(x,y):\forall i,t,\dot{x}_i(t)=\sum_j y_{ji}(t)-y_{ij}(t)} \int_0^T dt \sum_{i=1}^{k}\sum_{j\neq i} \left(\rho_{ij}(x(t))x_i(t) - y_{ij}(t)\right) + y_{ij}(t)\log\frac{y_{ij}(t)}{\rho_{ij}(x(t))x_i(t)} \end{aligned} \tag{43}$$

A.1.4 LDP for Population Processes Using Observables Alone

We can make some headway on this by using the calculus of variations. Specifically, we can turn the constrained minimization in Eq. 44 into an unconstrained problem by adding in Lagrange multipliers — we'll need one multiplier function $\lambda_i(t)$ for each type i, and they will (in general) need to be functions of time[15]. So we're looking for the unconstrained minimum of

$$\begin{aligned} & \int_0^T dt \mathcal{L}(x,\dot{x},y,\lambda) \\ & \equiv \int_0^T dt \sum_{i=1}^{k}\sum_{j\neq i} \left(\rho_{ij}(x(t))x_i(t) - y_{ij}(t)\right) + y_{ij}(t)\log\frac{y_{ij}(t)}{\rho_{ij}(x(t))x_i(t)} \\ & \quad - \int_0^T dt \sum_{i=1}^{k} dt\lambda_i(t)\left(\dot{x}_i(t) - \sum_{j\neq i} y_{ji}(t) - y_{ij}(t)\right) \end{aligned} \tag{44}$$

Now we invoke the general rule of the calculus of variations, that when we're extremizing an integral like this, where the inputs to the integrand are various functions $f_1, f_2, \ldots$, the

[15] One might think that there would need to be an additional constraint, to ensure that the trajectory stays on the simplex, that $\sum_i x_i(t) = 1$ for all t. But $\frac{d}{dt}\sum_i x_i(t) = \sum_i \dot{x}_i(t)$, and if the constraints on the $\dot{x}_i$s hold, that sum is necessarily 0 (because each y_{ij} appears in it twice, with opposite signs). Thus any trajectory which begins on the simplex, and obeys the constraints, will stay on the simplex automatically.

solution must obey the Euler-Lagrange equations

$$\frac{\partial \mathcal{L}}{\partial f_i} - \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{f}_i} = 0 \tag{45}$$

Here the functions in question are the x_i, the y_{ij}, and the λ_i.

This is easiest for the Lagrange multiplier functions, since they just give us back the constraint equations (as they are supposed to):

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}}{\partial \lambda_i} - \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{\lambda}_i} && (46)\\ &= \dot{x}_i(t) - \sum_{j\neq i} y_{ji}(t) - y_{ij}(t) && (47)\\ \dot{x}_i(t) &= \sum_{j\neq i} y_{ji}(t) - y_{ij}(t) && (48) \end{aligned}$$

It is also fairly straight-forward for the y_{ij} functions (remembering that each y_{ij} will show up in two constraint equations):

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}}{\partial y_{ij}} - \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{y}_{ij}} && (49)\\ &= \log\left(\frac{y_{ij}(t)}{\rho_{ij}(x(t))x_i(t)}\right) - \lambda_i(t) + \lambda_j(t) && (50)\\ y_{ij}(t) &= \rho_{ij}(x(t))x_i(t)e^{\lambda_i(t)-\lambda_j(t)} && (51) \end{aligned}$$

Finally, for the x_i functions, we get

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}}{\partial x_i} - \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{x}_i} && (52)\\ &= \sum_{p=1}^{k}\sum_{q\neq p}\left(\rho_{pq}(x(t))\delta_{ip} + x_p(t)\frac{\partial \rho_{pq}(x(t))}{\partial x_i}\right)\left(1 - \frac{y_{pq}(t)}{\rho_{pq}(x(t))x_p(t)}\right)\\ &\quad -\frac{d}{dt}(-\lambda_i(t)) && (53)\\ \dot{\lambda}_i(t) &= \sum_{p=1}^{k}\sum_{q\neq p}\left(\rho_{pq}(x(t))\delta_{ip} + x_p(t)\frac{\partial \rho_{pq}(x(t))}{\partial x_i}\right)\left(e^{\lambda_i(t)-\lambda_j(t)} - 1\right) && (54) \end{aligned}$$

where δ_{ip} is the Kronecker delta, which is 1 when the subscripts agree and 0 otherwise.

Substituting Eq. 51 into Eq. 48 gives

$$\dot{x}_i(t) = \sum_{j \neq i} e^{\lambda_j(t) - \lambda_i(t)} \rho_{ji}(x(t)) x_j(t) - e^{\lambda_i(t) - \lambda_j(t)} \rho_{ij}(x(t)) x_i(t) \tag{55}$$

The k functions on the left are fixed (since we are constraining the process to follow a particular trajectory of time x, which implies the derivatives $\dot{x}$), so we have a system of k equations in k unknowns, determining the λs.

We can also substitute in for the value of the rate function:

$$J(x) = \int_0^T dt \sum_{i=1}^{k} \sum_{j \neq i} \rho_{ij}(x(t)) x_i(t) \left(1 - e^{\lambda_i(t) - \lambda_j(t)} \left(1 - \lambda_i(t) + \lambda_j(t)\right)\right) \tag{56}$$

Since (as we've just seen) the λs are fixed by the trajectory x, this is a well-defined function of the trajectory, if somewhat less explicit than would be ideal.

To revert to the SI example from the warm-up, there was only one flux, in the present notation Y_{12}, so $y_{12}(t) = -\dot{x}_1(t) = x_2(t)$, and therefore $\lambda_1(t) - \lambda_2(t) = \log \dot{x}_2(t) / \rho_{12}(x(t)) x_1(t)$. Substituting into Eq. 56 does, indeed, give us back Eq. 28.

A.1.5 Limitations

Above, we have assumed that the transition rates ρ are functions of the instantaneous distribution over types $x(t)$. This assumption can be weakened somewhat; for example, ρ could be a function of the *history* of x up to and including time t, and nothing would change, *provided* the transitions are independent of the actual past trajectory given the rates ρ (i.e., provided ρ is a sufficient statistic for the transitions). As mentioned, models like this have often been employed in studying the self-organization of institutions [Young, 1998].

Perhaps more seriously, in making the transition rates functions *only* of $x(t)$ (or its history), we have imposed two restrictions. First, the absolute size of the population n is not allowed to affect the transition rates. Thus, one has the same rate of infection if 10 people are

infectious in a population of 20, as if a million people were infectious in a population of two million. This might be rationalized if we had the idea that each individual can only interact with so many others during each unit of time, and those "alters" (to use the sociological jargon) were themselves randomly drawn from the population. This however leads to the second limitation, which is that each individual is assumed to interact with the *aggregate* of the population; this corresponds to what epidemiologists call a "well-mixed population" (or "well-mixed compartment") and what physicists call a "mean-field approximation". Even if one wanted to say that, at each time step, an individual interacts with a random sample of available alters, and applied the ρ function to the distribution of that random sample, that would mean that each individual would have a *different* transition rate, which would itself be a random quantity. Our analysis above would then amount to replacing these random rates with their averages, neglecting the additional fluctuations which will result from the sampling[16]. There would also be the tricky question of whether each individual would sample its random alters *once*, at the beginning of the process, or would keep doing many random samplings, and, if so, how to make sense of that changing in continuous time[17]. If the number of alters is very large, the additional noise from their random choice is presumably small, and indeed subject to its *own* LDP, but it would need to be that carefully analyzed.

To put this limitation in perspective, however, we should point out that nothing prohibits having k being very large, say that each of c spatial sites or locations has its own version of the r different types (so $k = cr$). It would be possible, then, for the transition rate for one type at one site to be a function only of the distribution of types at that site, which would lead to r distinct copies of the original dynamics. If transition rates at one site are strongly dependent on the distribution within a site, and only weakly dependent on the distribution at other sites, we have a situation where there is mostly within-site interaction,

[16]In the social-learning case, the analysis would be appropriate if each agent played a game with *every* other member of the population, averaging the pay-offs. This should be close to what would happen if each agent played a game with a *large* random sample of other agents, but how close?

[17]In the physics jargon, the analysis, which replaces the random sampling by it average effect, is an "annealed" approximation, as opposed to a "quenched" one which would take the random sampling as done once and for all at the beginning.

complemented, perhaps, by migration from site to site. Nothing in the analysis would need to change. This device would *not* work, however, if the number of sites is to also grow with n. Thus we do not *directly* get results that apply to network models, where each individual interacts with a different set of individuals and so has (potentially) its own transition rates, since then the number of sites is just n.

A.1.6 Transitions between Stable States

Since institutional beliefs are self-reproducing, and, in a population sharing (roughly) the same institutional beliefs those beliefs reinforce each other, an institution is a stable state of the population. Transitions between institutions are therefore transitions between stable states. One advantage of the abstract, large-deviations approach is that it gives us some information about the time needed to move between stable states, and the paths by which such transitions happen.[18] Because of its importance for our argument, and because it shows something of the power of the large deviations approach, we sketch some of the main conclusions, with a very rough, heuristic derivation[19].

Suppose that a stochastic process obeys an LDP, with a rate function of the form of Eq. 3. Then, generically, the expected time it takes the process to move from one stable[20] state s to another state r will be exponential in $V(s,r)/\epsilon$, for a "potential" $V(s,r)$ that reflects the cost of the least-action path from state s to state r. Slightly more formally, define $\phi(s,r,T)$ to be the set of trajectories which start at s, and reach r for the first time at time T,

$$\phi(s,r,T) \equiv \{x : [0,T] \to \mathcal{X} \mid x(0) = s, x(T) = r, t < T \Rightarrow x(t) \neq r\} \tag{57}$$

Then in turn define $V(s,r,T)$ as the cost of the least-action trajectory from s to r in time

[18]The pioneering work on using large deviations to study transitions between stable states, "exit times" from the domain of a stable attractor, and transition paths, is that of Freidlin and Wentzell [1998]. The subject has been extensively treated in physics because of its importance for the "metastability" of some states of matter; see Olivieri and Vares [2005, §2.6, chs. 5, 6] for rigorous modern treatments and references, which include explicit calculations of the potential $V(s,r)$ (defined below) for many model systems.

[19]A slightly different, heuristic derivation goes back to **??**.

[20]That is, the state s would be stable in the limiting dynamical process when $\epsilon = 0$.

T,

$$V(s,r,T) = \inf_{x \in \phi(s,r,T)} \int_{t=0}^{T} L(x(t), \dot{x}(t)) dt \tag{58}$$

and define the potential as the cost of the least-action trajectory of any duration:

$$V(s,r) = \inf_{T>0} V(s,r,T) \tag{59}$$

Finally, define $\tau(s,r,\epsilon)$ to be the random amount of time it takes the process to move from state s to state r when the noise level is ϵ. Then, quite generically, the expected value of this exit time grows exponentially with the potential:

$$\lim_{\epsilon \to 0} \epsilon \log \mathbb{E}\left(\tau(s,r,\epsilon)\right) = V(s,r) \tag{60}$$

Because the initial state s is stable, in the limiting, $\epsilon = 0$ dynamics the process would never move from s to r, so the time needed would be infinite. In the presence of noise, however, if the transition is possible at all, it will happen eventually, and, conditional on its happening, it is exponentially more likely to happen along the least-action trajectory than along any other. To (roughly) see this, imagine writing the expected time as an integral:

$$\mathbb{E}\left(\tau(s,r,\epsilon)\right) = \int_{T=0}^{\infty} T \mathbb{P}\left(\tau(s,r,\epsilon) = T\right) dT \tag{61}$$

(slightly abusing notation to use $\mathbb{P}()$ as a probability density function). But, by their definitions,

$$\mathbb{P}\left(\tau(s,r,\epsilon) = T\right) = \mathbb{P}\left(X \in \phi(s,r,T)\right) \tag{62}$$

and the probability of the set on the right-hand side decays exponentially fast, following the large deviations principle. In fact, the minimum value of the rate function on $\phi(r,s,T)$ is

precisely $V(s, r, T)$. So

$$\begin{aligned} \mathbb{E}\left(\tau(s, r, \epsilon)\right) &= \int_{T=0}^{\infty} \int_{x \in \phi(s,r,t)} T\mathbb{P}\left(X = x\right) dx dT & (63) \\ &\approx \int_{T=0}^{\infty} T \exp\left\{-V(s, r, T)/\epsilon\right\} dT & (64) \\ & & (65) \end{aligned}$$

Laplace's method for approximating the integrals of exponential functions [Erdélyi, 1956, pp. 36-39] tells us that, as $\epsilon \to 0$, Eq. **??** approaches

$$\sigma T_0 \exp\left\{-V(s, r, T_0)/\epsilon\right\} \tag{66}$$

where T_0 is the minimizer of $V(s, r, T)$, and σ is a constant; also, $V(s, r, T_0) = V(s, r)$, by the latter's definition. So[21]

$$\epsilon \log \mathbb{E}\left(\tau(s, r, \epsilon)\right) \approx \epsilon \log \sigma + \epsilon \log T_0 + V(s, r) \to V(s, r) \tag{67}$$

Obviously, there are many steps here which would need to be made more precise in formal proofs. Sometimes this requires appealing to detailed facts about the rate function or the underlying process, but results of this form hold quite generally. One can also use such methods to extract more detailed information about the distribution of the transition time τ [Olivieri and Vares, 2005].

Notice that the time spent on the least-action trajectory from s to r, written T_0 above, just depends on the action function, and not on the noise level ϵ. The transition time scales exponentially with ϵ essentially because, at low noise, the process spends a lot of time wandering around the state s, before "finding" the path to r. Once embarked on that path, however, the actual transition can be very rapid.

Applied to evolutionary processes, this observation suggests that once a population has

[21] It is important here that $V(s, r) > 0$, so that the transition wouldn't happen in the absence of noise.

come to the neighborhood of a local fitness maximum or selective, self-reinforcing equilibrium, it will spend a long time there before it moves to the vicinity of another equilibrium, but the passage from one local maximum to another can be comparatively rapid. It was, indeed, a whole generation ago that Newman et al. [1985] argued, on exactly this basis, that "Neo-darwinian evolution implies puncutuated equilibria". Interesting questions arise as to whether such transitions happen primarily via crossing selectively-disfavored "barriers" or on the contrary along selectively-neutral paths, and in the barrier-crossing case, whether it is the height or the width of the barrier that determines the potential [van Nimwegen and Crutchfield, 2000], but this is secondary, for our purposes, to the fact of metastability [van Nimwegen et al., 1997]. Socially, of course, this is related to the proverbial way that institutions can seem eternal, only to collapse with stunning rapidity.

A.2 Network Structure and Contagions

[[Placeholder: Yank the most math-y material from §**??**]]

Move most math-y material to App. A.2]]. Plainly put, human societies tend to be characterized by gross asymmetries of power (although some societies are certainly more unequal than others), which manifests itself in patterns of social connectedness. Some individuals are highly influential - these individuals tend to be preferentially connected to each other, forming a relatively tight cluster. The theoretical and empirical literature suggests that this has important consequences for diffusion processes, since this cluster may be immensely influential in determining which contagions spread across the network, and which die out.

This literature identifies three important ways in which gross network topology affects contagion processes.First, the higher the average number of connections (social relationships along which beliefs can be transmitted) between nodes (individuals), the faster diffusion will be, and the less likely it will be to die out. Second, the higher the *transitivity* of the network (the more likely that if nodes A and B are both connected to C that A and B will also be connected to each other), the less likely it is that beliefs will spread across the network

as a whole. Networks with high transitivity tend to consist of tightly connected clusters of nodes, also called "communities" or "modules" which are much more weakly tied to other clusters. Equally, once a belief has appeared within a given cluster, it is more likely to spread and become established within that cluster.[22].The first node in a cluster to adopt the innovation, say A, finds many susceptible neighbors, and so initially the innovative can spread rapidly. However, when a node B adopts the innovation from A, transitivity implies that B's neighbors are also likely to be A's neighbors, and so no longer susceptible. After a fairly short amount of time, there are few susceptible nodes left within the cluster, and few links outside the cluster, so the innovation becomes, as it were, endemic but not pandemic. Finally, the higher the *variance* in the number of links associated with particular nodes (their "degree"), the easier it is for innovations to diffuse across the population. High variance implies that some nodes will have especially large numbers of links to other nodes, which may include non-local nodes. In networks with high variance in the number of links between nodes (e.g. networks with scale-free or lognormal distributions of links) the most linked nodes play a very important role in determining whether behaviors are disseminated or not. [23]

The action function not only reflects the mechanism of selection, but also the composition

[22]In fact, the most widely used methods for finding communities or clusters work by finding the borders across which diffusion is inhibited: Newman, 2006, Lee and Wasserman, 2010 [[TODO: Cite Lerman]]

[23]Since (as discussed below) ties tend to form between nodes with similar attributes, it is always possible that apparent diffusion across a social network is really due to similar individuals reacting similarly to a common external cause [Sperber, 1996]. Disentangling this from contagion or social influence raises tricky statistical issues, and often requires heroic assumptions [Shalizi and Thomas, 2011] #Cosma-add-cite-to-your-new paper? . But no one disputes that social influence exists.

The common element of results across a very wide range of these models is that evolutionary processes obey *large deviations principles*, under which the probability of a certain historical trajectory $x(t)$ of strategies in the population is roughly exponential, $\approx C \exp\left\{-\epsilon^{-1} \int_{t=0}^{T} L(x(t))dt\right\}$, the integral within the exponent being the "action" (or relative difficulty; low action trajectories are relatively easy to find, while high action trajectories are relatively hard) of the trajectory, and the scaling factor ϵ reflecting the noise level of the evolutionary process. Put this in Appendix I? ϵ reflects the consequences of the *mechanism of variation* for change across a given population. It grows with the frequency and magnitude of variations, and shrinks with the population size (the larger the population, the more difficult it will be for a given variation to spread across it).[24] The action function reflects the consequences of the mechanism of selection — but also, as already noted, the size of the selectively-neutral set of alternatives or "neutral network" that must be explored before a transition to a higher-fitness state [van Nimwegen and Crutchfield, 2000].

of the population of possible variations over which this mechanism selects. If there are many selectively-neutral variations, and few selectively-superior variations, then it will take longer on average to move to a higher fitness state than it would if there were fewer selectively-neutral variations and more selectively superior ones. If there is little 'noise', so that new variations are produced slowly, change from one state to another can take a very long time indeed.

TODO: replace with below text and move to Appendix I, integrated with math. The large deviations principle implies [Freidlin and Wentzell, 1998] that the typical time it takes a population

References

Daron Acemoglu and James A Robinson. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. Broadway Business, 2013.

Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000. doi: 10.1038/35019019. URL `http://arxiv.org/abs/cond-mat/0008064`.

R. R. Bahadur. *Some Limit Theorems in Statistics*. SIAM Press, Philadelphia, 1971.

Mary L. Boas. *Mathematical Methods in the Physical Sciences*. Wiley, New York, second edition, 1983.

Duncan S. Callaway, M. E. J. Newman, Steven H. Strogatz, and Duncan J. Watts. Network robustness and fragility: Percolation on random graphs. *Physical Review Letters*, 85:5468–5471, 2000. doi: 10.1103/PhysRevLett.85.5468. URL `http://arxiv.org/abs/cond-mat/0007300`.

Federica Carugati. *Creating a Constitution: Law, Democracy, and Growth in Ancient Athens*. Princeton University Press, Princeton, NJ, 2019.

Federica Carugati, Josiah Ober, and Barry Weingast. Is development uniquely modern? athens on the doorstep. 2015.

Elisabeth S Clemens and James M Cook. Politics and institutionalism: Explaining durability and change. *Annual Review of Sociology*, 25(1):441–466, 1999.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer Verlag, New York, second edition, 1998.

Frank den Hollander. *Large Deviations*. American Mathematical Society, Providence, Rhode Island, 2000.

Jeremiah E Dittmar and Skipper Seabold. Media, markets and institutional change: The protestant reformation. *Center for Economic Performance Discussion Paper, London, UK*, 2015.

Richard S. Ellis. *Entropy, Large Deviations, and Statistical Mechanics*. Springer-Verlag, Berlin, 1985.

Stephen P. Ellner and John Guckenheimer. *Dynamic Models in Biology*. Princeton University Press, Princeton, New Jersey, 2006.

A. Erdélyi. *Asymptotic Expansions*. Dover, New York, 1956.

Gregory L. Eyink. Action principle in nonequilbrium statistical dynamics. *Physical Review E*, 54:3419–3435, 1996. doi: 10.1103/PhysRevE.54.3419.

Henry Farrell. The consequences of the internet for politics. *Annual Review of Political Science*, 15, 2012.

M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer-Verlag, Berlin, second edition, 1998. First edition first published as *Fluktuatsii v dinamicheskikh sistemakh pod deistviem malykh sluchainykh vozmushchenii*, Moscow: Nauka, 1979.

Andrew Gelman. Letter to the editors regarding some papers of Dr. Satoshi Kanazawa. *Journal of Theoretical Biology*, 245:597–599, 2007. URL `http://www.stat.columbia.edu/~gelman/research/published/kanazawa.pdf`.

Avner D. Greif and David D. Laitin. A theory of endogenous institutional change. *American Political Science Review*, 98:633–652, 2004. doi: 10.1017/S0003055404041395. URL `http://ssrn.com/abstract=548363`.

Jacob S Hacker, Paul Pierson, and Kathleen Thelen. Drift and conversion: Hidden faces of institutional change.

Peter A. Hall and Kathleen Thelen. Institutional change in varieties of capitalism. *Socio-Economic Review*, 7(1):7, 2009. ISSN 1475-1461.

Jaako Hintikka. *Principles of Mathematics Revisited*. Cambridge University Press, Cambridge, England, 1996.

Ronald L Jepperson. The development and application of sociological neoinstitutionalism. In Joseph Berger and Morris Zelditch, editors, *New Directions in Contemporary Sociological Theory*, pages 229–266. Lanham, MD: Rowman & Littlefield, 2002.

Jack Knight. *Institutions and Social Conflict*. Cambridge University Press, Cambridge, England, 1992.

Jack Knight and Douglass North. Explaining Economic Change:¡ i¿ The Interplay Between Cognition and Institutions¡/i¿. *Legal Theory*, 3(03):211–226, 1997. ISSN 1469-8048.

Ann B. Lee and Larry Wasserman. Spectral connectivity analysis. *Journal of the American Statistical Association*, 105:1241–1255, 2010. doi: 10.1198/jasa.2010.tm09754. URL `http://arxiv.org/abs/0811.0121`.

Margaret Levi. Can nations succeed? *Perspectives on Politics*, 11(1):187–192, 2013.

Alfred J. Lotka. *Elements of Physical Biology*. Williams and Wilkins, Baltimore, Maryland, 1924. Reprinted as *Elements of Mathematical Biology*, New York: Dover Books, 1956.

John Stuart Mill. *Principles of Political Economy with some of their Applications to Social Philosophy*. Longmans, Green and Co., London, seventh edition, 1909. URL `http://www.econlib.org/library/Mill/mlPCover.html`.

Joel Mokyr. *A Culture of Growth: The Origins of the Modern Economy*. Princeton University Press, 2016.

C. M. Newman, J. E. Cohen, and C. Kipnis. Neo-darwinian evolution implies punctuated equilibria. *Nature*, 315:400–401, 1985. doi: 10.1038/315400a0.

Mark E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74:036104, 2006. URL `http://arxiv.org/abs/physics/0605087`.

Mark E. J. Newman and Juyong Park. Why social networks are different from other types of networks. *Physical Review E*, 68:036122, 2003. URL `http://arxiv.org/abs/cond-mat/0305612/`.

Douglass C. North. *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, Cambridge, England, 1990.

Douglass C. North, John Joseph Wallis, and Barry R. Weingast. *Violence and Social Orders: A Conceptual Framework for Interpreting Human History*. Cambridge University Press, Cambridge, England, 2009a.

Douglass C North, John Joseph Wallis, Barry R Weingast, et al. *Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History*. Cambridge University Press, 2009b.

Josiah Ober. *Democracy and knowledge: innovation and learning in classical Athens*. Princeton University Press, 2008. ISBN 0691133476.

Josiah Ober. *Mass and elite in democratic Athens: Rhetoric, ideology, and the power of the people*. Princeton University Press, 2009.

Josiah Ober and Barry R Weingast. The sparta game: Violence, proportionality, austerity, collapse. 2016.

Enzo Olivieri and Maria Eulália Vares. *Large Deviations and Metastability*. Cambridge University Press, Cambridge, England, 2005.

John F Padgett and Paul D McLean. Organizational invention and elite transformation:

The birth of partnership systems in renaissance florence1. *American Journal of Sociology*, 111(5):1463–1568, 2006.

John Frederick Padgett and Walter W Powell. *The Emergence of Organizations and Markets*. Princeton University Press, 2012.

Scott E. Page. *The Difference: How the Power of Diveristy Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press, Princeton, New Jersey, 2007.

Adam Przeworski. Institutions matter? *Government and Opposition*, 39(4):527–540, 2004.

Dani Rodrik. When ideas trump interests: Preferences, worldviews, and policy innovations. *Journal of Economic Perspectives*, 28(1):189–208, 2014.

William H. Sandholm. *Population Games and Evolutionary Dynamics*. MIT Press, Cambridge, Massachusetts, 2010.

Cosma Rohilla Shalizi and Andrew C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research*, 40:211–239, 2011. doi: 10.1177/0049124111404820. URL `http://arxiv.org/abs/1004.4704`.

Dan Sperber. *Explaining Culture: A Naturalistic Approach*. Basil Blackwell, Oxford, 1996.

Kim Sterelny. Cultural evolution in california and paris. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 62:42–50, 2017.

Hugo Touchette. The large deviations approach to statistical mechanics. *Physics Reports*, 478:1–69, 2009. doi: 10.1016/j.physrep.2009.05.002. URL `http://arxiv.org/abs/0804.0327`.

Erik van Nimwegen and James P. Crutchfield. Metastable evolutionary dynamics: Crossing fitness barriers or escaping via neutral paths? *Bulletin of Mathematical Biology*, 62: 799–848, 2000. URL `http://arxiv.org/abs/adap-org/9907002`.

Erik van Nimwegen, James P. Crutchfield, and Melanie Mitchell. Finite populations induce metastability in evolutionary search. *Physics Letters A*, 229:144–150, 1997. URL `http://www.cs.pdx.edu/~mm/fpimies.pdf`.

S. R. S. Varadhan. *Large Deviations and Applications*. SIAM, Philadelphia, 1984.

S. R. S. Varadhan. Large deviations. *Annals of Probability*, 36:397–419, 2008. doi: 10.1214/07-AOP348. URL `http://math.nyu.edu/faculty/varadhan/wald.pdf`. Wald Lecture, 2005.

H. Peyton Young. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton University Press, Princeton, 1998.