The evolution of patented inventions is Darwinian

Figures

Figures about a trait's heritability (transmission):

- Figure 1 – heritability/transmission parent vs. offspring where traits are keyword tfidf scores

- Figure 2 – heritability/transmission in actual parent-offspring pairs vs. randomly sampled pairs

- Figure 3 – heritability/transmission z vs. z' where traits are strength of LDA topic

- Figure 4 – heritability/transmission z vs. z'. Left: boolean trait z. Right: continuous trait z. Above: traits are tfidf of keywords. Below: traits are LDA topic strength.

- Figure 5 – expected (average) frequencies of boolean z = background trait frequencies $f(z)$. DRAFTED with Excel.

Figures about the connection between an entity's trait value and its fitness:

- Figure 6 – narrow fitness as a function of trait value.

- Figure 7 – wide fitness as a function of trait value.

- Figure 8 – narrow fitness as a function of trait deviation from the norm.

- Figure 9 – wide fitness as a function of trait deviation from the norm.

Trait lineage figures:

- Figure 10 – Diagrams some actual trait lineages. Small atlas of different kinds.

- Figure 11 – Ranked distribution of actual trait lineage lengths. Label top 10 or so. Left: traits = tfidf keywords. Right: traits = topic strength)

- Figure 12 – ranked distribution of trait lineage lengths given *pure drift*

Study *models*. Models of evolving*patent citation networks* (our previous work) vs. models of evolving *traits (information)* in patents patent citation networks.

- Make a *good* model of the evolution of traits/information in patented inventions. The model's fitness function needs to change over time, driven in part by exogenously evolving demands for different kinds of inventions. The *skeptic's hypothesis* is that no adequate model is simpler than the data itself, i.e., the data is algorithmically random (weakly emergent). A good model would falsify the skeptic's hypothesis.

- Confirm Paris results on evolution of patent *citation networks*: uniform, nonuniform but fixed, preferential attachment, preferential attachment and aging (Valverde). Modify nonuniform model with a knob that tunes model behavior from the target to full neutrality. One example: Include a knob for the widow size, that controls how quickly the fixed non-uniform fitness function changes over time. Generate the initial population of patents with a nonuniform fitness function measured empirically over an initial time window. The minimal knob setting reproduces the actual patent record (statistically, over many trials), and the maximal knob setting is the fixed non-uniform "a priori" model from Paris.

- We assume pure drift as a null hypothesis, and study models of drift, and examine deviation from those models. Make a neutral model of evolution of *traits (information)* flowing through a patent citation network. XXX

- Compare actual trait lineages with lineages in data created by *resampling, reshuffling* minibatches (time chunks) of target data. Reshuffling the data over time destroys many evolutionary patterns through successive minibatches.

- Create a random corpus by randomly sampling the empirical distribution of word frequencies). Then measure tfidf keywords and LDA topics in this random corpus. Test whether keywords and topics in target corpora are significantly different from "random" keywords and topics.

1   Introduction

We ask whether the evolution of patented inventions is Darwinian. This is equivalent to asking whether the population of patented inventions evolves by the Darwinian process of natural selection. Many different kinds of populations are Darwinian; natural selection most obviously shapes populations of biological organisms, but is also shapes populations of other kinds of entities, such as genes, groups of organisms (e.g., altruists vs. cooperators), and cells in certain tissues (e.g., cancerous tumors). We ask whether natural selection shapes one specific further population: patented inventions.

Most treatments of cultural evolution (CITE Cavalli-Sforza and Feldman, Boyd and Richerson, Mesoudi) treat culture as the beliefs, norms, institutions, etc. of populations of *humans*. These treatments explain the evolution of human culturally transmitted traits by a variety of processes, including Darwinian natural selection and non-Darwinian cultural drift, and they focus on the interaction between cultural and biological evolution.

Our approach is different. We foreground human *artifacts*, like Brian Arthur (CITE), and put humans in the background. Our specific focus is the population of *patented inventions*. Patented inventions directly embody a part of human culture. Our approach measures how natural selection shapes the transmission of traits into new patented inventions.

Of course, patented inventions exist only because of humans. Inventions are created through a complex social and cultural process that involves a network of interactions among a great many people, including inventors, designers, producers, consumers, users, suppliers, lawyers, patent examiners, among others. All of these people affect the evolution of patented inventions. At the same time, those humans are themselves affected by many non-human environmental factors, such as economic conditions, supply of components and other resources, technological facility, weather, among many other things. We view humans as part of the complex social and cultural background environment that shapes (and is shaped by) the evolution of technology. In the foreground are the cultural artifacts themselves.

We study patented inventions by means of a proxy: the first page of the invention's patent record, which contains the patent's title, abstract, issue date, and prior art (i.e., the earlier patents it cites), among other things. The patent record contains virtually complete, accurate, and uniform information about every single patented invention. This makes patents an especially promising context for empirically documenting Darwinism shaping cultural artifacts.

2   Darwinian evolution XXX fit with GPE perspective on Darwinism!

We call a population (set) P of entities *Darwinian* if and only if there is a set of traits or *information elements*, X, and each entity $i \in P$ has a subset of those traits, X_i, and:

1. *Information elements vary* across the entities in the population; typically, $X_i \neq X_j$ if $i \neq j$. At any given time typically the distribution across the population of frequencies of information elements is non-uniform. Furthermore the distribution changes over time. This change in the distribution of the population of information element frequencies is the evolution of the information in the population of entities.

2. *Information elements are transmitted to new entities* in P from existing entities in P. The production process maps individual information senders to information receivers, and it maps individual information elements sent to elements received. The ancestral (recursive chain) of (sender,receiver) pairs defines both the entire *genealogy of entities* in P and the entire *genealogy of information elements* in X. The information transmitted down lineages can be affected by a variety of parameters, including the rates of mutation, recombination, migration, and drift, among others.

3. *Information elements are transmitted into new entities at different rates*, so the information elements have differential fitness values. The fitness at a given time of a given information element in a given entity depends on the entity's selective environment at that time, which might include the other entities in the population (as in frequency-dependent selection). The fitness of a given information element typically changes over time, because its selective environment changes.

If a population meets these Darwinian conditions, then it will evolve by natural selection. Other processes like drift or migration might also have an affect, but natural selection will be a significant factor. If the evolution of patented technology is Darwinian, this tells us something about what drives traits into new patented inventions.

Mesoudi (CITE) distinguishes Darwinian evolution (i.e., evolution in a Darwinian population) from *Neo-Darwinian* evolution, which is the special form of Darwinian evolution with three features that were discovered after Darwin to be important in biological populations. Specifically, Neo-Darwinian evolution is defined as evolution a Darwinian population in which traits are discrete rather than continuous, in which acquired traits are not inherited (no Lamarkianism), and in which genetic innovation like mutation is unaffected by intelligence, intention, or affect on fitness. A Neo-Darwinian population is a Darwinian population, by definition, but some Darwinian populations are *not* Neo-Darwinian. For example, if cultural traits are continuous rather than discrete, if acquired cultural traits are inherited, and if new cultural traits are produced by conscious human intention and intelligence, as Mesoudi argues, then cultural evolution is *not* Neo-Darwinian. Nevertheless, if humans and their culturally transmitted traits meet all the conditions of Darwinian populations, as Mesoudi also argues, the evolution of human cultural traits would be Darwinian.

2.1 Darwinism among patented inventions

To determine whether the evolution of patented inventions is Darwinian we simply check for the properties listed above. We think of the traits of a patent as a small subset of information elements. Patents contain widely varying information, although it is possible for two different patents to contain exactly the same information. And new patents are produced by a process that depends on certain existing patents, which are recorded as the patent's *prior art*. So, patented inventions have some hallmarks of Darwinian populations. Whether they have *all* the hallmarks and are a Darwinian population depends on whether two remaining conditions obtain:

1. Information elements are *transmitted* from senders to receivers. In particular, they are transmitted at a significantly higher rate between a patent and its prior art than between random pairs of patents. The key is that the information transmission process is high enough that it significantly affects the evolution of the distribution of trait frequencies.

2. Information elements are *transmitted at different rates* to receivers, and thus their fitness values differ. In some populations there is a positive expected background rate of information transmission between random pairs of patents, even in the absence of selection (i.e., under pure drift). In such populations, an empirical sign of Darwinian evolution is when information transmission significantly exceeds background transmission.

If the evolution of patented inventions meets these two final conditions, then information transmission rates can be explained by Darwinian natural selection among the population of new patents.

2.2   Cumulative Darwinian evolution

Darwinian evolution of traits in a population is said to be *cumulative* if it consists of a sequence of individual episodes of natural selection that produce complex adaptations. The idea is that a complex adaptation is built up step by step in a sequence of episodes of natural selection. Each separate episode creates and incorporates a separate component of the complex adaptation, and these components incrementally accumulate in increasingly complex adaptations through a sequence of episodes of natural selection. The final complex adaptation exists only when the sequence of episodes ends.

A Darwinian population undergoes natural selection, so it will produce adaptations. But those adaptations might all be equally simple. We consider only traits that are individual information elements, Z; all such traits are equally simple. The Darwinian evolution of equally simple traits could never be cumulative. This paper focuses only on whether evolution is Darwinian.

ACTION ITEM: Figure out how to *measure* cumulative evolution in the patents. Need degrees of simple vs. complex traits, accumulate complexity incrementally. See Lenski et al. Nature paper. Also Brian Arthur. Perhaps *combinations* of information elements

Darwin himself and many others (e.g., Peter Godfrey-Smith) believe that cumulative Darwinian evolution is the most plausible explanation for complex biological adaptations found in nature. Some (e.g., Mesoudi) believe that cumulative Darwinian evolution similarly explains the complex human cultural traits. By contrast, Brian Arthur complexity growth in technologies using a process that is not even Darwinian, much less not cumulative.

Does the complexity of patent traits grow over time, in general? The simple answer is that we don't know. All patented inventions must be judged by patent examiners to be *novel*, so perhaps the average amount of the new information in patents has remained relatively constant over time. Perhaps highly fit information does not accumulate in patents. (Did someone recently claim that the number of innovative patents (= those without any prior art) remained constant in time??)

Nevertheless, technological artifacts ("widgets") do contain complex innovations. Furthermore, complex traits accumulate by combining separate innovations together. The iPhone, for example, contains many patented inventions. Also it seems that there is an arrow of complexity in technology that is analogous to the arrow in biology; that is, the complexity of the most complex widgets seems generally to be growing over time. It follows that many individual patents are useful only in a larger human and technological environment in which a number of other patents are also useful.

3   Formalism for multisexual evolution of patented technology

Production of new entities in *asexual* populations is especially simple, because each offspring has exactly one parent. (We ignore horizontal information transfer for the moment.) *Bisexual* offspring production is a little more complicated, because each offspring has exactly two parents. In *multisexual* populations an offspring has a multiplicity of parents (typically more than two but fewer than twenty), and different offspring in the same population might have different numbers of parents. The population of patented technology is multisexual. Most mathematical formalisms for evolution assume that the population is asexual or bisexual but cultural evolution is typically multisexual. Hence, we use a general mathematical formalism for evolution in multisexual populations developed by to Kerr and Godfrey-Smith (CITE).

   We consider an episode of evolution which happens from time t^a to time t_d. The entities in the population O at time t_d are the *descendants* of the entities in population P at time t^a, which are their *potential ancestors*. We let n^a be the size of P and n_d be the size of O. In the present context, the transition from t^a to t_d corresponds to an episode in which a "ring" of new growth (O) is added to the tree trunk of existing patents (P).

3.1   The matrix of citations (productions), C

The notation for an evolving multisexual population is build from a *non-sparse* $n^a \times n_d$ boolean matrix, C, which records whether or not source patent i is cited by receiver patent j:

$$C_j^i = \begin{cases} 1 & \text{j cites i} \\ 0 & \text{otherwise} \end{cases}$$

(Note that the indices i and j range over patents (entities) in populations P and O.) The matrix C is an indicator function filled with 1's and 0's that records which multisexual parent-offspring interactions produce O from P. We now define some further concepts using C:

$C_*^i = \sum_{j=1}^{n_d} C_j^i$: the "hit count" of i (during the episode from t^a to t_d), i.e., the number of descendant entities in O which cite ancestor i, i.e., the number of receivers for source i. The hit count of an ancestral entity i is also interpreted as its *(realized) fitness* (see paragraph below).

$C_j^* = \sum_{i=1}^{n^a} C_j^i$: the "prior art count" of j (during the episode from t^a to t_d), i.e., the number of ancestral entities in P which cite descendant j, i.e., the number of sources for receiver j.

$C_*^* = \sum_{i=1}^{n^a} \sum_{j=1}^{n_d} C_j^i$: the total number of citations (production connections) between ancestral entities at t^a and descendant entities at t_d, i.e., the number of ancestral entities in P which cite some descendant in O, i.e., the number of (source, receiver) pairs (i, j), i.e., the total number of production connections between populations P and O.

C_*^a : the hit count (total number of citations received) of "an ancestor" (a variable, with instances like C_*^i). This 1d vector is created by collecting counts in the 2d matrix C across all columns (descendants).

C_d^* : the prior art count (total number of citations made) of "a descendant" (another variable). Created by collecting counts in C across all rows (ancestors).

$\tilde{C}^a_* = C^a_*/(C^*_*/n^a)$: the *relative* hit count (number of citations received), relative to other patents in P.

$\tilde{C}'^*_d = C^*_d/(C^*_*/n_d)$: the *relative* prior art count (number of citations made), relative to other patents in O.

The concept of fitness becomes is more subtle in bisexual than asexual populations, and more subtle still in multisexual populations. The fitness of an entity is typically measured by counting the number of offspring produced by the entity. But a multisexual population is *multi-parental*, and this dilutes a parent's opportunity to transmit traits to its offspring. It turns out that the fundamental concept of fitness can be defined as C^i_*, the hit count of i. This is the *(realized) fitness* of ancestor i, i.e., the number of *descendants* attributable (at least partly) to ancestor i.

Descendants in multisexual populations are produced by (i.e., cite) different numbers of ancestors (different amounts of prior art), so the amount of fitness accredited to an ancestor (the hit count fraction) of a given descendant is diluted by the credit due to the other ancestors (the rest of the prior art). So, a more fine-grained measure of the fitness of multisexual ancestors would account for this dilution. We define the *diluted (realized) fitness* of i, C^i_*/C^*_j as the sum of descendant fractions attributable to i:

$$\widehat{C}^i_* = \sum_{j=1}^{n_d} \left(C^i_j \frac{1}{C^*_j} \right),$$

where the *descendant fraction* of j that is attributable to i is defined as the reciprocal of C^i_* (and where this reciprocal is defined as equal to 0 if $C^i_j = 0$). The corresponding expression for diluted relative fitness is

$$\widetilde{\widehat{C}^i_*} = \frac{\widehat{C}^i_*}{\frac{1}{n^a} \sum_{i=1}^{n^a} \widehat{C}^i_*}.$$

3.2   The trait, X

We let X be the value of some measurable characteristic or trait of entities in P and O. For example, the trait X might be the $tfidf$ score of a keyword in a patent. A given patent has many keywords, so it has many traits, and X is just one of those traits. Our treatment of one trait, X, can be duplicated for any finite number of other measurable characteristics.

We define the following quantities concerning the traits in ancestral and descendant populations:

X^i : the value of trait X in ancestral entity $i \in P$ (instance).

X_j : the value of trait X in descendant entity $j \in O$ (instance).

X^a : the "ancestral character" (variable).

X_d : the "descendant character" (variable)

$\overline{X^a} = \frac{1}{n^a} \sum_{i=1}^{n^a} X^i$: the average character value in the ancestor population P.

$\overline{X_d} = \frac{1}{n_d} \sum_{j=1}^{n_d} X_j$: the average character value in the descendant population O.

7

Evolution or change of the trait X (ΔX) is measured as the change in average character value between the ancestral population P at time t^a and the descendant population O at time t_d:

$$\Delta \bar{X} = \overline{X_d} - \overline{X^a}.$$

We can also measure the change in X value between a specific ancestor, i, and one of its descendants, j, as follows:

$$\Delta X_j^i = X_j - X^i.$$

We also define the *average change* in X from sources to receivers:

$$\text{ave}(\Delta X_d^a) = \frac{1}{C_*^*} \sum_{i=1}^{n^a} \sum_{j=1}^{n_d} (C_j^i \Delta X_j^i) = \frac{1}{C_*^*} \sum_{i=1}^{n^a} \sum_{j=1}^{n_d} C_j^i (X_j - X^i)$$

where the indicator function C is used to find the (sender, receiver) pairs in the non-sparse matrix C. We also define two variables concerning the covariance between an entity's trait values and the count of its hits or prior art:

$\mathbf{cov}(\tilde{C}_*^a, X^a)$: the covariance between an ancestor's X value and its relative hit count.

$\mathbf{cov}(\tilde{C}_d^*, X_d)$: the covariance between a descendants X value and the relative prior art count.

The covariances and averages are defined over the entire population $P \cup O$, rather than random samples from the population, in the standard way (e.g., Wikipedia), for real-valued random variables X and Y:

$$\text{cov}(X, Y) = \text{ave}[(X - \text{ave}(X))(Y - \text{ave}(Y))]$$
$$= \text{ave}(XY) - \text{ave}(X)\text{ave}(Y)$$

3.3   Generalized Price Equation (GPE)

One of the attractions of the *Generalized Price Equation* (or GPE) derived by Kerr and Godfrey-Smith is the ability to handle descendants variable ancestry. The population of patents contains descendants with variable numbers of ancestors (patents which cite different amounts of prior art). Evolution in populations with variable number of ancestors is termed *multisexual*. Multisexual evolution differs from bisexual evolution, and both differ from asexual evolution. Different modes of reproduction can produce different kinds of evolution. Parallel with the standard "ancestral" point of view of Price's equation, an ancestral formulation of GPE is as follows:

$$\Delta \bar{X} = \frac{\text{cov}(C_*^a, X^a) + \text{ave}(C_*^a (\Delta X)^a) - (n_d/n^a)\text{cov}(C_d^*, X_d)}{C_*^*/n^a}, \tag{1}$$

where

$$(\Delta X)^i = \frac{\sum_{j=1}^{n_d} C_j^i (X_j - X^i)}{C_*^i}$$

is the "average deviation in character between ancestral entity i and its connected descendant entities" (and is interpreted as equal to 0 if the ancestral entity is unconnected). The first two terms

correspond to the terms in the original Price equation, and the third term is "like a mirror image of the first term."

The GPE need not be expressed from the "ancestral" perspective shown in Equation 1. Kerr and Godfrey-Smith also express GPE from a more neutral, a purely connection-based perspective, as follows:

$$\Delta \bar{X} = \text{cov}(\tilde{C}_*^a, X^a) + \text{ave}(\Delta X_d^a) - \text{cov}(\tilde{C}_d^*, X_d). \tag{2}$$

This discussion will use the connection-based form of GPE expressed in Equation 2.

The GPE partitions evolutionary change into three components, corresponding to the three terms in Equation 2:

- $\text{cov}(\tilde{C}_*^a, X^a)$ reflects *differential fecundity* (differential "divergence") of trait values. This term is often described as reflecting the strength of Darwinian natural selection.

- $\text{ave}(\Delta X_d^a)$ reflects *production bias* (differential "transformation") towards certain trait values, e.g., mutation or inheritance biased toward higher X values.

- $\text{cov}(\tilde{C}_d^*, X_d)$ reflects *differential "convergence"* and it takes different forms. One form is *differential innovation*, e.g., mutation that is biased to certain trait values. Another form is *differential dilution* of ancestors with certain trait values; e.g., mixture of asexual, bisexual, and multisexual modes of reproduction.

3.4 Categories of evolutionary change

A simple pattern among the three GPE components can be used to define a "fingerprint" of certain different kinds of evolutionary change. GPE fingerprints are codes consisting of bit strings, where each bit encodes the zero or non-zero value of a GPE term in their order shown in Equation 2. E.g., GPE code `010` corresponds to the category of evolutionary change in which $\text{cov}(\tilde{C}_*^a, X^a) = 0$, $\text{ave}(\Delta X_d^a) \neq 0$, and $\text{cov}(\tilde{C}_d^*, X_d) = 0$. The $2^3 = 8$ different GPE codes (`010`, `001`, ...) signify eight different basic categories of evolutionary change. The GPE codes sum to one if the evolutionary change involves only one GPE component, i.e., if only one GPE term is non-zero. The "impure" categories of evolutionary change that mix different pure categories have GPE codes that sum to two or three. (The GPE code of no evolutionary change sums to zero.) Table 1 lists the GPE fingerprint and statistical signature of the some of the categories of evolutionary change.

The GPE provides a specific perspective on how to decide if an evolving system is undergoing Darwinian natural selection. Many people equate "Darwinian" evolution with selection for fecundity. If the pure form of Darwinian evolution has GPE code `100`, then the impure forms of Darwinian evolution have forms `101,110`, and `111`. Pure and impure forms of Darwinian evolution can all be compactly represented by adding wildcards ("don't care") in the GPE fingerprint: `1**`. So, the question whether the evolution of patented technology is Darwinian can be empirically answered by examining whether the evolution depicted in the patent record has this *Darwinian GPE fingerprint*.

We use the Generalized Price Equation to decompose evolutionary change over some period of time into three components and calculate its GPE fingerprint. Our first task is to measure the GPE fingerprint over the entire time spanned by our corpus. This will indicate the overall category of evolutionary change occurring. Next we can measure the GPE fingerprints in the patent record

GPE code	$\text{cov}(\tilde{C}^a_*, X^a)$	$\text{ave}(\Delta X^a_d)$	$\text{cov}(\tilde{C}^*_d, X_d)$	Category of evolutionary change
100	non-zero	zero	zero	Pure differential fecundity. KGS Fig. 2A.
1⋆⋆	non-zero	(any)	(any)	Pure or impure differential fecundity.
010	zero	non-zero	zero	Pure *production bias*. KGS Figure 2B.
⋆1⋆	(any)	non-zero	(any)	Pure or impure production bias.
001	zero	zero	non-zero	Pure *differential innovation*. E.g., migration. KGS Figure 2C.
001	zero	zero	non-zero	Pure *differential dilution*. KGS Fig. 2D.
⋆⋆1	(any)	(any)	non-zero	Pure or impure differential innovation or dilution.
110	non-zero	non-zero	(any)	Pure differential fecundity and production bias.
11⋆	non-zero	non-zero	(any)	Pure or impure differential fecundity and production bias.
101	non-zero	(any)	non-zero	Pure differential fecundity and differential innovation or dilution.
1⋆1	non-zero	(any)	non-zero	Pure or impure differential fecundity and differential innovation or dilution.
011	(any)	non-zero	non-zero	Pure production bias and differential innovation or dilution.
⋆11	(any)	non-zero	non-zero	Pure or impure production bias and differential innovation or dilution.
111	non-zero	non-zero	non-zero	Differential fecundity, production bias, and differential innovation or dilution.
000	zero	zero	zero	No evolutionary change.

Table 1: Some categories of evolutionary change defined as combinations of the processes associated with each GPE term.

over a sequence of shorter time periods, e.g., a year, and then see if and how the category of evolutionary change exhibited by the system changes over time.

4 Results

4.1 Heritability: (X^a, X_b)

We measure *heritability* of trait X as the slope of the regression line in the scatter plot of X_d (descendant trait value) as a function of X^a (ancestral trait value). We can measure this slope even if X is a boolean-valued trait (present or absent).

Figure 1: Scatterplot of points (X^i, X_j), for (ancestor, descendant) pairs $i \in P$ and $j \in O$, with a linear regression line. Here, the continuously valued trait X is *TFIDF* score of a term (e.g., high-content words like 'stent', 'PCR', 'non woven web', compared with low-content words like 'invention', 'part', 'mechanism').

Figure 2: Compare heritability (transmission) of trait X in (ancestor, descendant) pairs compared with random pairs of patents. Left: (X^i, X_j) where (i, j) is an (ancestor, descendant) pairs as in Figure 1. Right: (X^i, X_j) where (i, j) are random pairs of patents $i \in P$ and $j \in O$.

Figure 3: Scatterplot of points (X^i, X_j), for (ancestor, descendant) pairs $i \in P$ and $j \in O$, with a linear regression line, like Figure 1 except that the trait X is the strength of an *LDA* topic.

4.2 Fitness and deviation from fitness norm XXX work in progress!

Recall the distinction between *(realized) fitness* and *diluted (realized) fitness*. We want to determine to what extent the value of the trait X^i of ancestor i is correlated with i fitness (C_*^i) and diluted fitness C_*^i / C_j^*.

Figure 4: Comparison of boolean vs. continuous traits X: scatterplots of points (X^i, X_j), and regression lines, where trait X is a term's *tfidf* score (for terms like 'stent' vs. 'invention'). Left: X is a boolean trait, valued as 1 or 0 depending on whether or not the *tfidf* score of 'stent' exceeds a threshold determined by hand in advance. Right: X varies continuously with *tfidf* score.
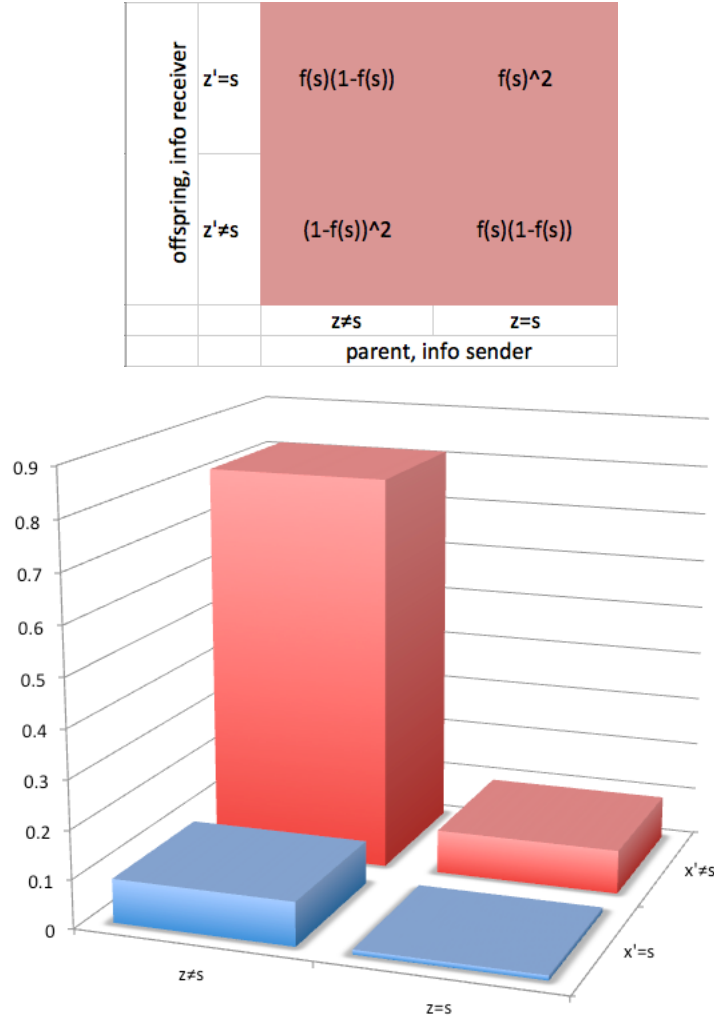

offspring, info receiver		$z \neq s$	$z = s$
	$z'=s$	f(s)(1-f(s))	f(s)^2
	$z' \neq s$	(1-f(s))^2	f(s)(1-f(s))
		parent, info sender	

Figure 5: Expected frequencies, $E(X_d)$, of boolean traits $z_i \in \{0, 1\}$ (e.g., significant keyword 'stent') given only background word frequencies, $f(X)$. In this case, word frequencies are unaffected by selection, drift, mutation, migration, etc. Background keyword frequencies, $f(X)$, can be empirically measured for all traits z by measuring relative word frequencies of words in all the documents in the entire corpus. Above: Expected values, $\mathrm{E}(X^i, X_j)$, of boolean trait X given only background word frequencies, $f(X)$. Below: Expected values of trait X, given (an artificially high) background frequency $f(X) = 0.1$ of the trait X.

13

Figure 6: Scatterplots of pairs (X^i, w^i) of narrow (realized) fitness w^i of ancestor $i \in P$ as a function of trait X^i of i, showing how well trait values are correlated with narrow (realized) fitness. Traits X include high-content words like 'stent', 'PCR', and 'inkjet', and low-content words (in the context of patents) like 'invention', 'part', 'process', 'machine'.

Figure 7: Scatterplots of pairs (X^i, W^i) of wide (realized) fitness W^i of ancestor $i \in P$ as a function of trait X^i, showing how well trait values are correlated with wide (realized) fitness. We compare trait values, z_i, with squared deviations from trait averages, $z_i - \bar{z})^2$, and we compare narrow w_i and wide W_i fitness, for all (or representative sample of) $i \in P$. Traits X include high-content words like 'stent', 'PCR', and 'inkjet', and low-content words (in the context of patents) like 'invention', 'part', 'process', 'machine'.

Figure 8: Scatterplots of pairs $(z_i - \bar{z})^2, w_i)$ of narrow (realized) fitness w^i as a function of trait X^i, showing how well trait values of i are correlated with narrow (realized) fitness of i. Traits X include high-content words like 'stent', 'PCR', and 'inkjet', and low-content words (in the context of patents) like 'invention', 'part', 'process', 'machine'.

Figure 9: Scatterplots of pairs $(X_i - X)^2, W_i)$ of wide (realized) fitness W^i of ancestor $i \in P$ as a function of the squared deviation of i's trait X^i from the average ancestral value, $\overline{X^a}$, showing how well i's trait deviations from the norm are correlated with narrow (realized) fitness of i. Traits X include high-content words like 'stent', 'PCR', and 'inkjet', and low-content words (in the context of patents) like 'invention', 'part', 'process', 'machine'

4.3 Length of lineages of traits: NOTE over many time steps

We consider boolean-valued traits $z_i \in \{0, 1\}$ in patents. For simplicity, assume that 'stent' is the one trait, z. This trait is either present or absent in each patent. A boolean trait might be created by thresholding an underlying quantitative trait.

Trait lineages start whenever an offspring contains a trait that is in none of its parents (none of its prior art). A trait lineage contains a (parent,offspring) link whenever both patents contain the trait. A trait lineage is then the ancestral of the (parent,offspring) relation that follows from the introduction of a new trait in an offspring (new patent). The *length* of a lineage of is the number of patents linked together in the longest branch of offspring that descends after the introduction of a new trait.

Figure 10: The lineage of keyword 'stent', and lineages of keywords frequently co-occurring with 'stent'. Trait lineages overlay the lineages of patent parents and offspring. Time increases from top to bottom. Include a selective atlas of lineages of other keywords: 'invention', 'PCR', 'zeolite', 'computer', etc.

Figure 11: Length of lineage of boolean traits z as a function of ranked (descending) length of lineage, where traits are keywords with tfidf scores above a handpicked threshold (threshold = XXX). Left: traits z are words with *high* specific content in patents ('stent', 'inkjet', 'PRC', 'zeolite'). Right: z for words with *low* specific content in patents ('invention', 'machine', 'process').

Figure 12: Length of lineages produced in a *pure drift* null model of with no natural selection; other details as in Figure 11.
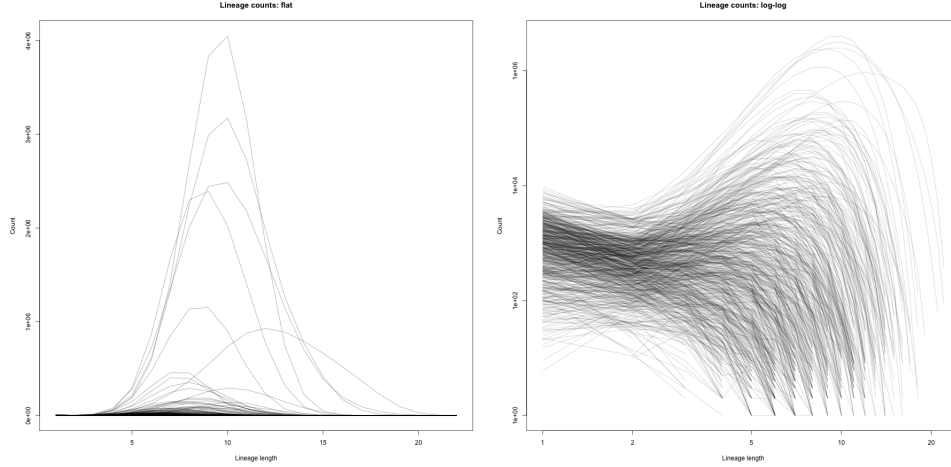
Figure 13: Distribution of counts of lineage lengths (for different words?), as a function of lineage length. On right: log-scale. Old fig from Devin.

5   NP take on Price Equation

I just spent the weekend camping with four papers, one by Price, three by Frank: Frank (1995, 1997, 2012); Price (1995). I read most thoroughly the 2004 review by Frank, as I found it to be the most useful overview. Here I give my thoughts from that relatively clean slate, which is to say, I did not take this manuscript to try and merge what I learned from the papers into Mark's current framework. I will now describe a fairly simple version of how we might interpret our data in the Price framework, and I look forward to elucidating the connections with Mark's work in the previous section.

Price's equation takes various forms... e.g.

$$\begin{aligned}
\Delta \bar{z} &= \bar{z}' - \bar{z} \\
&= \sum q_i' z_i' - \sum q_i z_i \\
&= \sum q_i' (z_i' - z_i) + \sum q_i' z_i - \sum q_i z_i \\
&= \sum (\Delta q_i) z_i + \sum q_i' (\Delta z_i)
\end{aligned}$$

Another version is:

$$\bar{w} \Delta \bar{z} = \mathrm{Cov}(w, z) + E(w \Delta z),$$

with

$$\mathrm{Cov}(w, z)/\bar{w} = \sum q_i (\frac{w_i}{\bar{w}} - 1) z_i = \sum (\Delta q_i) z_i,$$

(since $q_i' = q_i w_i / \bar{w}$) and finally, another useful version:

$$\Delta \bar{z} = \mathrm{Cov}(w, z')/\bar{w} + E(\Delta z).$$

These are all equivalent versions of Price's equation, and the details of their equivalence and their various merits are discussed by Frank (2012).

The basic interpretation is that

$$\Delta \bar{z} = \text{Cov}(w, z')/\bar{w} + E(\Delta z)$$
$$= \Delta_s(\bar{z}) + \Delta_E(\bar{z}),$$

where $\Delta_s(\bar{z})$ is the component of $\Delta \bar{z}$ that is due to *selection*, and $\Delta_E \bar{z}$ is what Fisher called 'the environmental component' of evolutionary change. Frank is careful to point out that really the 'environmental component' simply has every other form of evolutionary change that is not simply due to selection. Our real concern is whether the selection component, $\Delta_s(\bar{z})$, is large with respect to the other component.

I ended up thinking of i as labeling traits; Frank uses examples of alleles, genotypes, etc., but I think our traits are best thought of as grams or topics. [I think this is the first and maybe largest departure with Mark's formulation above, where i labels patents or 'parent sets' of patents.] A distinction is made between entities in the population "having a trait" (signified by the frequencies q_i), and "measuring a property of the trait" over the population (signified by the z_i). A key point: i is labeling traits, not individuals in the population. Frank rarely explicitly indexes members of the population; he simply refers to 'population averages'. q_i is the frequency, i.e. probability of the ith trait occurring in the population, i.e. n_i/N, where n_i is the number of entities with the trait and N is the total number of entities in the population.

Another key notational issue is related to the distinction between the 'ancestral population' and the 'descendant population'. Notationally, the quantities measured on the ancestral population are un-primed, and quantities measured on the descendant population are primed. But... all quantities subscripted with i are referring to the *ancestral* population. In particular, Frank states "... if $i = 2$ specifies a particular phenotype, then q_2' is not the frequency of the phenotype $i = 2$ among the descendants. Rather, it is the fraction of descendants derived from entities with the phenotype $i = 2$ in the ancestors." This is a tricky but essential point.

So... I am thinking that q_i measures the fraction of the patent population that contains a certain keyword (i.e. the ith keyword), and z_i is the average TF-IDF score of that keyword for that patent (averaged over the population). Or, i could label a topic, and z_i would be the average strength of that topic. The topics are a bit subtle; in the virtual Bayes framework used by gensim, all topics have some probability of being appropriate for any particular document, i.e. for any document the topic probability is greater than zero. Pragmatically, topic's are deemed relevant for a document of the topic probability for that document exceeds some threshold (e.g. 0.01 is the threshold used for gensim to print out a topic).

Sticking with the idea of keywords as traits and TF-IDF scores for measurements for simplicity, we would have q_i = the frequency of ancestral patents that contain keyword i; $q_i = n_i/N$, where N is the total number of ancestral patents. q_i' are the frequency of descendant patents descended from those ancestors containing keyword i, i.e. $q_i' = n_i'/N'$, where N' is the size of the descendant population and where the n_i' are computed by identifying the citation links from the n_i' descendants to the n_i ancestors.

An issue to resolve is definition of ancestral and descendant populations. The most natural analog to the biological discussion is to have a temporal separation, e.g. ancestral is all patents before the year 2000, and descendant is the patents in 2000. But patents in 2000 actually reference patents in 2000, so there is some 'reproduction' going on among the descendants, but I don't see anything wrong with that formally. The descendants could be taken to be all patents in 2000 that cite patents before 2000.

We have discussed having the parent population be all patents (in our database) that have ever been cited, and all descendants as all patents that have cited any patent in the database. Obviously a huge overlap in these two sets, but I'm not sure there is any formal difficulty in computing the elements of Price's equation. I'm not sure this characterization makes sense, however. Besides being highly dis-analogous with the biological context, it seems to lose a sense of evolution *through time*. I favor the temporal distinction, maybe with a shorter chunking than one year. For short enough time chunking, we eliminate reproduction within the descendent patent pool, as they will all be referencing only the ancestral pool.

Anyway, once we make a choice for ancestral and descendant patent populations, I think that given the citation links, all the relevant quantities are easily computable, for putting numbers to each of the terms in Price's equation. Unless I 'm missing something(which may be!), we can then observe empirically if the selection component, $\Delta_s(\bar{z})$, is large with respect to the 'environmental' component. If so, then we have a good case for Darwinian evolution.

[a side issue for us: is there any place in the literature that identifies Darwinian evolution with a relatively large $\Delta_s(\bar{z})$ in the Price equation? how about a treatment of drift vis a vis the Price equation? worth a google...]

One last comment for now: Frank mentions a paper, Kerr and Godfrey-Smith (2009), that specifically addresses innovation, i.e. introduction of new elements that are not descended from existing elements. He mentions "their formulation depends on making explicit the connection number between each individual ancestor and each individual descendant, rather than using the fitnesses of types..." This actually sounds like a formulation that might be appropriate for our patents. And maybe akin to Mark's more complex formulation. I have located the paper and put it in the dropbox, but I haven't read it yet.

6 Evolution of information in patented inventions XXX fit to GPE!

6.1 Population of patented inventions (entities)

We assume that there is a population of patented inventions, $P = \{p_1, p_2, ..., p_N\}$. P is the set of patents not that *could* be invented but that actually *are* invented. P grows over time as new patents are produced.

P_t denotes the set of patents that exist at t, and $P_{1:t}$ denotes the set of all patents generated through time t,

$$P_{1:t} = \bigcup_{1 \leq t' \leq t} P_{t'}. \tag{3}$$

The growth of the patent population is analogous to the process that adds new rings to a growing tree trunk. New patents are generated in the trunk's outermost ring and select information from sources deeper in the trunk. This analogy explains one of the ways in which patents differ from biological populations: the fact that patents never "die." Once a patent has been generated in the new ring around the tree trunk, that patent and all of the other patents in that ring continue to exist, along with every other previously issued patents. Thus, $P_t = P_{1:t}$, for all t. So, the new patents added at t, denoted P_t^{new}, are those existing patents that did not exist before:

$$P_t^{\text{new}} = P_t \setminus P_{t-1}, \tag{4}$$

where $A \setminus B$ denotes set difference.

6.2 Information elements in patents (traits)

We assume that there is a set, $Z = \{z_1, z_2, ..., z_V\}$, that contains all *possible* traits or information elements that any patent could ever have. For simplicity, we assume that we know in advance a fixed finite number V of possible information elements, analogous to a known dictionary of words in a corpus. Our guiding examples of patent traits are tfidf keywords and LDA topics, but our formalism is general and supports many other interpretations.

We consider only "elementary" information, which consists of sets of information elements. A more general approach would also consider "compound" information items composed by applying boolean operations (union, intersection, complement) to information elements and compounds.

The traits in patent r are a set of information elements, $Z_r = \{z_{r1}, z_{r2}, ..., z_{rn}\} \subset Z$. The number of traits, n, which is typically very small, corresponds to the number of "loci" in r. Note that different patents can have different numbers of information elements, so it is possible that $\#Z_i \neq \#Z_j$ if $i \neq j$. This is one more way in which the population of patents has mathematical properties that differ from the biological populations in which evolution is normally studied.

Information elements $z \in Z$ can be boolean variables, reflecting traits that are either entirely absent or present (have the value 0 or 1). Information elements $z \in Z$ can also reflect traits that are *weighted* if they are interpreted as scalar variables (or, equivalently, as a scalar weight paired with each boolean trait).

We use Shannon entropy, H, to quantify the amount of information in various trait distributions. For example, if we consider the information in patent r, Z_r, and let $Z_{rj} = z_{rj} \in Z_r$, then the

Shannon entropy of Z_r (in bits) is:

$$H(Z_r) = - \sum_{j=1}^{\#Z_r} Z_{rj} \log_2 Z_{rj}. \tag{5}$$

If the information elements are boolean, then the *amount of information* (in bits) in patent r simplifies to the log of the number of information elements in r, $\log(\#Z_r)$. Shannon entropy analogously quantifies the amount of information in various other trait distributions discussed below.

The *information pool* of n patents $p_1, p_2, \ldots p_n$ is the union of the information in each, $Z_1 \cup Z_2, \ldots \cup Z_n$. The *information overlap* of n patents $p_1, p_2, \ldots p_n$, is the intersection of the information in each, $Z_1 \cap Z_2, \ldots \cap Z_n$. The *information pool* in the entire population at time t, P_t, is the information in the patents that exist at t:

$$Z_t = \bigcup_{p_i \in P_t} Z_i. \tag{6}$$

The *historical information record* at t is the pool of information that has existed up to time t:

$$Z_{1:t} = \bigcup_{1 \le t' \le t} Z_{t'} = \bigcup_{p_i \in P_{1:t}} Z_i \tag{7}$$

Since the population of patents continually grows, the pool of information in the patent corpus at a given time t is equal to the union of the information in all of the patents generated up to that time, $Z_t = Z_{1:t}$. We measure the amount of information in any of these distributions of traits as its Shannon entropy: $H(Z_t), H(Z_{1:t}), H(Z_t^{\text{new}})$.

6.3   Information transmission to patents (reproduction and inheritance)

information in r and s overlaps: their intersection is nonempty, $Z_s \cap Z_r \ne \emptyset$

the information overlapping r and s: the intersection of their information, $Z_s \cap Z_r$

s transmits z to r: r cites s and their overlapping information contains z, $z \in Z_r \cap Z_s$

s transmits information to r: r cites s and their information overlaps, $Z_r \cap Z_s \ne \emptyset$.

   s sends information to r

   r receives information from s, r inherits information from s

the information s transmits to r: their overlapping information, $Z_s \cap Z_r$, given that r cites s

s is the *unique* source of information in r: some of the overlapping information of s and r is in none of r's other sources

Note that many patent information sources are not unique sources of any information in any patent. This is one of the many ways in which patented inventions differ from familiar biological populations. An information element can be transmitted to a patent by many different sources.

Patents $s_1 \ldots s_n$ are all sources of information element z in a patent, r, if r cites $s_1 \ldots s_n$ and z is in their overlapping information, $z \in Z_r \cap Z_{s_1} \cap \cdots \cap Z_{s_n}$.

Each new patent r produced at t appears in P_t^{new}. Generating a patent r involves generating its information, Z_r, and its *prior art* or information sources, $\{s : r \text{ cites } s\}$. Given the information in a patent, one can calculate the *information potentially transmitted* to patent r, denoted Z_r^{sources}, as the union of the information in all of r's sources:

$$Z_r^{\text{sources}} = \bigcup_{p_i \in \{s: r \text{ cites } s\}} Z_i. \tag{8}$$

The information *transmitted* to r, denoted Z_r^{trans}, is the overlap between the information in r and the pool of potentially transmitted information in r's sources:

$$Z_r^{\text{trans}} = Z_r \cap Z_r^{\text{sources}} \tag{9}$$

As usual, Shannon entropy measures the amount of information in these distributions of traits.

6.4 New information (innovation)

Even though every patented invention is certified as "novel" by a patent examiner, the information in one patent might be the same as the information in other patent; it is possible that $Z_i = Z_j$ for $i \neq j$. The new information in the pool at t, denoted Z_t^{new}, is the total information in the pool at t, Z_t, minus the previous information in the pool, Z_{t-1}:

$$Z_t^{\text{new}} = Z_t \setminus Z_{t-1}. \tag{10}$$

Sometimes patent contain information that appears in none of their sources. This is the analog of a "mutation" in the population of patents. The information in patent r can be partitioned into two subsets, the information transmitted from r's sources and the "new" information in r, $Z_r = Z_r^{\text{new}} \cup Z_r^{\text{trans}}$. Equivalently, *new information* in r, denoted Z_r^{new}, is the total information in r minus the information transmitted to r:

$$Z_r^{\text{new}} = Z_r \setminus Z_r^{\text{trans}}. \tag{11}$$

If a patent, r, is generated at some time, t, information that is new in r would still exist elsewhere in the patent population at t by being in patents that are not r's sources. In this case $Z_r^{\text{new}} \subset Z^t$, so the information that is "new" in patent r is *not* new in the whole population.

The information that appears in any of the new patents generated at t is simply the pool of their information. From this pool subtract the information that existed previously, to yield the new information elements that appear for the first time in any patent at t:

$$Z_t^{\text{new}} = \left(\bigcup_{p_i \in P_t^{\text{new}}} Z_i \right) \setminus Z_{t-1}. \tag{12}$$

As usual, we measure the amount of information in these distributions as Shannon entropy.

There are two further ways in which a patent's information might be unique and differ from all other patents. A patent r has a *new combination* of information at time t just in case $Z_r \neq Z_i$ for all $p_i \neq s$ such that $p_i \in P_t$. By contrast, a patent r has a *new element* of information just in case there exists an information element $z \in Z_r$ such that for all $p_i \neq r$, $z \notin Z_i$. A patent with a new information element also has a new combination of information, but a new combination of information might contain no new information elements.

6.5 Measuring information mixing (blending, diluting)

In asexual Darwinian populations with no horizontal gene transmission, new entities are produced when an existing entity buds or fissions, so information in new entities comes from one source. By contrast, in *sexual* Darwinian populations new entities mix information from multiple sources (typically two, one for each sex).

There is a worry that if new entities mix information from too many sources, then Darwinian natural selection will not explain the existence and spread of traits. XXX QQ - Is this a worry about natural selection or about cumulative selection specifically? XXX

We can measure a number of things about the extent to which the information in a recipient *mixes* information from many sources.

the mixing coefficient of r: the number of sources cited by r, $\#\{s : r \text{ cites } s\}$.

> XXX We want to calibrate this mixing coefficient so that standard sexual biological populations having a mixing coefficient of 2. In a sexual population without mutation, a child is expected to inherit roughly half its genes from each parent. So, we define the mixing coefficient to be the reciprocal of the expected average inheritance from each individual source ??? XXX

size of r's pool of potential source information. the size at t of the union of all the information in any of r's sources: $\#(\bigcup_{s \in P_t, r \text{ cites } s} Z_s)$. This is the pool of information from which a new patent selects which existing information to include (by selecting which existing patents to cite).

A number of analogous concepts can be defined, such as a patent's minimal number of sources, and its number of disjoint sources.

6.6 Measuring information transmission (inheritance)

We observe baseline frequencies of information elements (traits), $z \in Z_{1:t}$, in the historical record of information up to time t. We equate (interpret) these observed historical frequencies as the background unconditional probabilities (expected frequency) of information elements in the entire corpus:

$$\frac{\#\{i : p_i \in P_t \wedge z \in Z_i\}}{\#\{i : p_i \in P_t\}} \tag{13}$$

$$\approx \frac{\text{prob}(z \in Z_i \wedge p_i \in P_t)}{\text{prob}(p_i \in P_t)} \tag{14}$$

$$= \text{prob}(z \in Z_i \mid p_i \in P_t). \tag{15}$$

The rate at which information element $z \in Z$ is transmitted from source s to receiver r at time t ...

$$\frac{\text{prob}(z \in Z_r \mid r \text{ cites } s \wedge z \in Z_s \wedge r \in P_{1:t})}{\text{prob}(z \in Z_s)}. \tag{16}$$

The *transmission rate* of information element $z \in Z$ from source s to receiver r, can be measured as the fraction (percentage) of (s,r) pairs in which z is transmitted from s to r:

$$\frac{\#\{(s,r) : r \text{ cites } s \wedge z \in Z_s \cap Z_r\}}{\#\{(s,r) : r \text{ cites } s \wedge z \in Z_s\}}. \tag{17}$$

This is the fraction of *possible transmissions* of z that are *actual transmissions*.

The background observed frequency of z in Z_t makes us expect that there is some nonzero probability that (frequency with which) possible z transmissions actually happen. We measure the background expected frequency of transmission between in random pairs of patents by choosing pairs of patents at random with uniform probability, and observing the rate at which information would have been transmitted if one had cited another:

$$\text{prob}(z \in Z_r | z \in Z_s \wedge r \in P \wedge s \in P). \tag{18}$$

6.7 Rate of information flow into new patents (fitness of a trait)

Recall that P is the set of all patents in an entire corpus, and P_t is the set of new patents generated at time t. One can think of new patents as "selecting" which existing patents will serve as their information sources, and thereby "selecting" which information elements are in the pool that could be transmitted to new patents. Information elements "compete" for selection as information sources for new patents.

The entities in Darwinian populations by definition have *traits* with differential fitnesses. The fitness of a trait is relative to the environment in which the trait is expressed, and a traits fitness might change if its environment changes. A highly fit trait will tend to spread through the population of new entities more quickly than expected given just that trait's present frequency in the corpus. Our reference corpus at t is the set of all patents that exist up to t, $P_{1:t}$.

We would like to measure the extent to which term frequencies (in a mini batch) are higher than would otherwise be expected given merely background term frequencies. E.g., let's consider the fitness at t of the keyword $z = $ *stent*.

We define the *fitness* of information element z at the t relative to reference corpus $P_{1:t}$, as the fraction (percentage) of *new* patents in the mini batch, Q_t, that contain the information, z, compared to the fraction of patents in the whole reference corpus that contain z:

$$\frac{\#\{i : z \in Z_i \wedge i \in P_t^{\text{new}}\}/\#P_t^{\text{new}}}{\#\{i : z \in Z_i \wedge i \in P_{1:t}\}/\#P_{1:t}}. \tag{19}$$

XXX OR: the fitness of an information trait $z = $ the probability that a patent is cited (up to time t) given that it contains information element z XXX

6.8 Sources driving information flow (fitness of an entity)

The entities in Darwinian populations also have differential fitness values. The fitness of an *entity* in a Darwinian population in general is the extent to which the entity's traits (information) influence the production of the same traits in new entities. We define the *fitness* of an individual patent i at time t as simply its citation rate at t [XXX is this a rate? XXX]:

$$\#\{j : j \text{ cites } i \wedge i \in P_t\}. \tag{20}$$

23

This is the number of new citations i receives at t, that is to say, the number of i's opportunities to be an information source for another patent at t. The total number of a patent's opportunities to be an information source by time t, or its *cumulative fitness*, is $\#\{j : j \text{ cites } i \ \wedge i \in P_{1:t}\}$.

7 Discussion and conclusions

References

Frank, S. A. (1997). The Price equation, Fisher's fundamental theorem, kin selection, and causal analysis. *Evolution; international journal of organic evolution*, pages 1712–1729.

Frank, S. A. S. (1995). *George Price's contributions to evolutionary genetics*, volume 175. Journal of Theoretical Biology.

Frank, S. A. S. (2012). Natural selection. IV. The Price equation. *Journal of Evolutionary Biology*, 25(6):1002–1019.

Kerr, B. and Godfrey-Smith, P. (2009). Generalization of the Price equation for evolutionary change. *Evolution; international journal of organic evolution*, 63(2):531–536.

Price, G. R. (1995). The nature of selection. (Written circa 1971, published posthumously). *Journal of Theoretical Biology*, 175(3):389–396.