BRILL

Journal of Cognition and Culture 13 (2013) 287–311

brill.com/jocc

Positive Frequency Dependence in Graffiti: An Empirical Case Study of Cultural Evolution

Julia A. Maddison and Michael Doebeli*
Department of Zoology, University of British Columbia, 4200–6270
University Boulevard, Vancouver, BC, Canada V6T 1Z4
* Corresponding author, e-mail: doebeli@zoology.ubc.ca

Abstract
Cultural traits can be seen to evolve by a process similar to natural selection. They are transmittable, variable, and the variants can have differential fitness. As a result, cultural evolution can in principle lead to non-random distribution of cultural traits. A limited number of studies have addressed the evolution of human cultural traits "in the wild," partly because culture is difficult to categorize into discrete units. Parallel to studying non-random species distributions in ecosystems due to natural selection, we have focused on investigating non-random distributions of cultural traits in a local environment. We used a collection of library study desks to categorize graffiti into content-based cultural traits, or "topics", and quantified the level of clustering for each topic as a measure of non-random distributions of topics on the desks. Clustering was found to occur for some topics but not others, and the level of clustering varied with topic in ways that are consistent with topic content characteristics.

Keywords
Cultural evolution, cultural traits, natural selection, graffiti, empirical, clustering

Introduction

Background

Biology has found an extremely fruitful paradigm in the theory of evolution by natural selection. Many researchers have found parallels between the evolution of genetic material and the change over time of human culture, and we will briefly explain the basic principles. Reviewing the criteria for natural selection, it is clear that some elements of culture have the necessary characteristics of heritable variation and differential fitness associated with that variation. We will call selection on cultural traits "cultural selection", following Rogers and Ehrlich (2008) in distinguishing natural selection, which affects gene frequencies, from selection that affects frequencies of cultural traits.

 DOI: 10.1163/15685373-12342095

First, human culture is extremely diverse, with different subjects and different types of cultural traits. Second, it is replicated (transmitted) from peer to peer as well as from parents to children. While this transmission is not always accurate, in some cases, for example specifically phrased jokes or catch phrases, the fidelity in heredity can be high. In theoretical cases it has been shown that accurate transmission is not necessary for cultural traits to undergo cultural selection (Henrich and Boyd, 2002).

Third, the extent to which this replication occurs can depend on the inherent properties that vary between different cultural traits, i.e., there is differential fitness. One example comes from urban legends, whose replication success can depend on the emotions evoked by specific legends or versions of the same legend (Heath et al., 2001). Heath et al. (2001) show that versions of urban legends which participants found less credible were also considered more likely to be told again because of the possibility of gaining a higher emotional response. Further, in non-human cultural traits, it has been demonstrated that the functionality of communication affects what sorts of sounds spread in bird song in different regions (Cardoso and Atwell, 2010) and that the learned ability to acquire food in experimental conditions can be transmitted over multiple simulated generations of chimpanzees. (Horner et al., 2006; Mesoudi and Whiten, 2008). Interaction between the cultural trait and chimpanzee psychology create a selective environment in which the trait powerfully sweeps through the groups in a diffusion chain model (Horner et al., 2006). There are cultural traits that spread because of their inherent qualities as opposed to lucky placement in history, or drift in small populations. Thus, it appears that many cultural traits under reasonable conditions can be said to undergo cultural selection.

Because cultural selection can be expected to generate non-random distributions of cultural traits, we addressed the possibility of cultural selection in graffiti by comparing observed distributions of graffiti to randomly generated control distributions. To do this, we first had to develop phenotypic classes of graffiti on library desks, and then measure and evaluate the distribution of these phenotypes across desks. Clustering would reflect an association between cultural units with the same phenotype. This could be caused by various mechanisms, either due to interactions between similar cultural units, or due to external pressures which affect similar cultural traits in the same way.

Objectives

Our first objective was to find a phenotypic characterization of the cultural units, which involved semantically categorizing topics of graffiti markings and

exploring the methodological difficulties therein. The purpose of this objective is to acknowledge the difficulties in categorizing cultural traits because of the subjectiveness of such a categorization, and to explore possible solutions to this problem. We use the words "categories" and "topics" interchangeably, both referring to the semantic categories which represent different cultural traits or phenotypes. Our second objective was to observe whether clustering is occurring in the populations of graffiti for some or all of the semantic categories developed. We did this by collecting information about the frequencies of each category in each population (population = desk) and determining whether the variance in these frequencies across populations is high relative to the variance distribution produced by a simulation of randomly distributing all markings across all desks. A high variance in the frequencies across populations would suggest clustering, whereas a random distribution of the markings across desks represents a situation with no association between markings on desks, and thus serves as a null hypothesis.

Methods

Study Area

The study area is on the first floor of Koerner Library, at the University of British Columbia (UBC), Vancouver, Canada. The University of British Columbia has more than 45,000 undergraduate and graduate students. In total, there are approximately 729,000 visits to UBC's libraries each year, an average of 1998 per day (UBC Library Services Statistics, 2010). Given that all days are not equal, this value is probably much higher during exam season and much lower during the slow summer months (Margaret Friesen, Assessment Librarian, personal communication, 2011). About 10,000 unique students are estimated to use Koerner Library specifically each year, most of them visiting during the fall and winter sessions (M. Friesen, pers. commun., 2011).

There are 188 individual desks on the first floor of Koerner library that are known to have not been cleaned of graffiti for at least three years (K. Madill, pers. commun., 2010). The majority of these desks are small, wooden, cubicle style desks of the same format. All desks included in the study are cubicle style with wood panels separating desks from each other (Fig. 1), and only those that had not been cleaned were included.

Desks are spaced in clustered arrangements that are spread relatively evenly across the first floor of Koerner. Desks are arranged into clusters along the edges of circular pillars, in blocks of four to eight, or in pairs along walls (Fig. 1).

Figure 1. Arrangement of desks in blocks of six and around pillars (in background).

Data Collection and Categorization

Collection of Data. Every uncleaned desk was given a number, and 50 of these desks were chosen randomly for analysis. Of these, desks with fewer than 10 markings were removed from the study, yielding 31 desks to be sampled. For the 31 desks, all graffiti was recorded by hand. Three broad types of graffiti were encountered: writing in English, writing in another language, and doodles or drawings (Fig. 2).

Writing in English was transcribed into a notebook, taking note of whether markings were part of a conversation thread based on arrows and spatial location. The majority of such conversations were clearly distinguishable because arrows had been drawn to imply response to a given marking. Writing in another language was recorded based on which general language group the marking fell into, such as "East Asian writing." The classification of language had no later consequence on results, as anything in a language other than English was considered to be "unclassifiable" in terms of topic. Doodles or drawings were reproduced into a notebook if possible (e.g. a heart or a smiley face), or if too complex or unrecognizable (detailed geometric design), digital photos were taken for later consideration during categorization.

Development of Categories. Out of the 31 desks, 9 were chosen at random to develop the most common topics that would be used to categorize markings in the other 22 desks. Every marking from these 9 desks was pooled into one

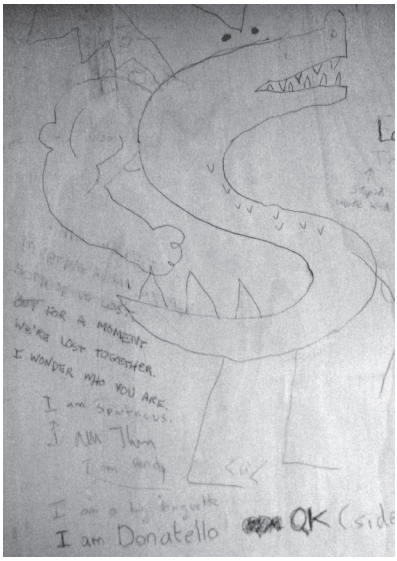

Figure 2. Example of graffiti.

spreadsheet and randomly sorted. For each marking, for 30 seconds, one of us (JAM) then performed a word association test by writing down important nouns and other related words that the marking evoked, limiting every marking to 30 seconds of this association. These tests were performed with some constraints in mind: the characterizing words had to be nouns, or very relevant adjectives, and could not be general statements about the kind of statement or thought-process someone might have had when writing the marking. For example, a marking like "JS <3 TR" would be characterized by the words "heart," "love," "romance," "relationship," but not the words "cute," or "cheesy," because the latter words describe a judgement of the marking, not its content. Upon developing a list of words for each marking, each of these words was placed into a broad category, such as "sports" for the words "game" or "baseball," or "sexgen" for anything about gender, intercourse, or genitalia. Then the number of times a given category appeared was counted. The word association depended in part on JAM's personal ideas about connotations, and her knowledge of various subjects. However, the goal of this step was to discover which topics were most common across all markings, in order to be sure of having a large enough sample size within each topic. As such, finding robust categories is more important than finding the absolute biggest category, and this process served to find relatively common categories. It also served, via the word association test, to develop keywords for each category and illuminate the potential difficulties and blurred boundaries inherent in certain categories.

Topic Definitions and Labels

Based on the graffiti from the 9 "calibration" desks chosen for defining the topics, the topics chosen for analysing the remaining 22 desks were "school", "sex/gender", "insult", "love/romance", "religion/cognition", "race/national" and then the "everything else" category, "unclassifiable." The reason for labelling some of these categories with two words is that it helps pinpoint the specific bent of each category. For example, "race" might imply only markings about Caucasians, African Americans, Mongoloids, etc., whereas including "national" allows the category to imply Koreans (Table 1).

Categorizing Markings

With the topics chosen and the markings recorded, each marking from the 22 desks that were not part of the calibration exercise was characterized by assigning it a binary string of 0s and 1s of length seven, according to whether it belonged to one or more of the seven topics. Specifically, each marking was given a 1 for a topic if the marking was considered to fall into that topic, and a 0 otherwise. The same marking could be in multiple categories, and hence have more than one 1's in its binary string. If a marking was in none of the six specific topics, the marking was placed in "unclassifiable" category. A marking could not be unclassifiable in addition to being in any other topic. During classification JAM noted whether she was uncertain about a given categorization, in order to later come back and develop deciding rules that would be consistent across all markings. JAM then pooled all of the uncertain markings, identifying common situations of uncertainty. For example, should "I love him but can't say" fall under both love/romance and sex/gender or just love/romance? By the rules developed (Table 1), any comments which were already in love/romance and contained a gendered pronoun were also classified as sex/gender. This was because in these scenarios, the gendered pronouns carried an important implication. The specific choice of rules was subjective and somewhat arbitrary, but the existence of a rule-based classification was necessary to remove some of the bias and discrepancy that JAM's specific subjectivity would induce. With explicit rules, the bias is at least repeatable by another researcher.

Splitting of Conversations

Markings that were clearly arranged as a conversation with multiple correspondents were initially considered together as a single, whole units. This data set was called "integrated," in which each conversation counted as one marking. A second data set was obtained by splitting conversations into individual responses, i.e., by considering each marking belonging to a conversation as a separate unit. This data set was called "separated."

Table 1
Definitions of categories, including keywords and explanations for ambiguous cases.

Topic	Keywords and key phrases	Examples
School	Class, study, school, essay, homework, test, exam, quiz, stressed, procrastinate.	"I hate studying" "I'm so grateful for this study space"
Sex/Gender	Anything about gender, intercourse, or genitalia. him, her, sex, bitch, cock, etc.	"I love boobs" "All men aren't assholes"
Insult	You suck, – suck(s), suck it up, you're stupid, you can't spell, you're gay.	"Suck it up princess" "Stupid white kid"
Love/Romance	<3, love, crush, in love, falling in love, brokenhearted, jealous.	"I love DY" "When will you marry me?"
Religion/Cog	Life, God, Jesus, Heaven, Hell, religion, philosophy, Jews, Muslim, Allah.	"God is great" "Love is all"
Race/Nat	Racist, race, racism, Asian, Indian Chinese, black (given context), brown, yellow, Caucasian, Canadian, American, etc.	"Do Asian girls like white guys?" "Everyone likes brown guys"
Unclassifiable	Markings that fit into no other category get a "1" under unclassifiable.	"Audi A4181" "The cake is a lie"

Variance Analysis

The goal of the variance analysis was to determine if there was clustering of specific topics on individual desks, i.e., in single graffiti populations. For the analysis we focused on each topic separately. For example, desk i had a certain number $M(i)$ of markings in the topic school, and a certain number $N(i)$ that were not. Then $f(i)=M(i)/(M(i)+N(i))$ is the frequency of markings on desk i that belong to the topic school. We then calculated the variance in the frequencies $f(i)$ over all desks, which contains information about the distribution of the topic school across all the desks. In particular, if the topic school were strongly clustered, some desks would have a frequency $f(i)$ close to zero, and the remaining desks would have a frequency $f(i)$ close to 1. Therefore the variance in the relative frequencies across all desks would be high. On the other hand, if the topic school were hyperdispersed, then all desks would have similar $f(i)$ values, and hence the variance in the $f(i)$ would be low. The same reasoning holds for the variance of the frequencies of all other topics across the desks.

In order to know how high the variance must be to indicate strong clustering, we performed randomization to develop a null distribution of variances. Each marking has either a 1 or a 0 for school, yielding total number of 0's and 1's for this topic. By redistributing these 1s and 0s randomly across all desks, while maintaining each individual desk's population size, we effectively mimic a situation in which there is no selective pressure for any association (or disassociation) of markings in this topic. For any given random distribution, we then calculate the various frequencies $f(i)$ across all desks, as well as the corresponding variance in the $f(i)$. In this way, each random distribution yields a variance, and we repeated this process 1,000,000 times to generate a null distribution of variances. The variance calculated from the unaltered data can be compared to this null distribution to obtain a p-value by finding the percentage of variances in the null distribution that are equal to or greater than the estimated variance. For example, if 7% of the variances of the null distribution were equal to or greater than the observed variance, the p-value is 0.07, which is the probability that the observed value of the variance has occurred by random chance. Thus, if the p-value is small and the observed variance is higher than most variances in the null distribution, this is an indication that some non-random process led to clustering of a topic. A Java program was written to perform the simulations, run on MRJ version 1060.1.6.0_20-279 on a Mac OS X 10.6.4 computer. The accuracy of the p-values was checked by randomizations in JMP and Open Office (SAS Institute, 2009; OpenOffice.org, 2010).

Topic Correlation

The variance analysis addresses whether or not individual topics are clustered across desks, while the topic correlation analysis begins to address whether two topics are correlated across desks. To do so, we compared the relative topic frequencies from the above analysis between each pair of topics across all desks. In other words, we tested whether the frequency of a given topic correlates with the frequency of another topic across all desks. We used JMP to test for correlation between the frequencies (SAS Institute, 2009). Because topic frequencies are non-normal distributions, we used the non- parametric test of Spearman's ρ.

Categories Survey

The final component of this study was a survey designed to test the robustness of our categorization of markings into topics. The survey was a selection of 29 markings that university students in three classes then categorized into the seven topics in the same way that we classified them. Two of the classes had one set of 29 markings, the other class had a different selection of markings with some overlap. The overlap was intended to qualitatively reveal any potential patterns at the class level that would reflect a relevant difference in survey instructions or in student base.

To be clear, the survey was not randomized among students or in terms of the markings chosen. The students were in the same classes with JAM. The markings with the most uncertain or ambiguous classification were chosen, and as such we expected some disagreement between students' classification. Markings were also selected to be as minimally offensive as possible given the variety of markings available. The purpose of the survey was to begin to understand the limitations and strengths of our categorization method, especially to understand what sort of rules need to be made in order to make our method repeatable.

Percentage agreement was calculated for each marking within each topic, and was initially separated by class. The percentage values reflect the percentage of students who gave the majority classification, such that if most students gave a particular marking a 0 for a given topic, the percentage value reflects how strong that majority is. We compared the percentage agreement of the students with our own classifications, focusing on instances in which the majority of students agreed with our classification, and the instances in which the majority of students disagreed with our classification.

Results

Graffiti General Characteristics

The number of markings on desks ranged from 13 to 88 in the integrated data set (conversations treated as single units), with an average of 31.6±19.3 (Mean and SD) markings per desk, and from 19 to 141 markings per desk in the separated data set (conversations treated as multiple units), with an average of 46.6±30.6 markings per desk (Fig. 3). The number of markings that were categorized in a given topic, excluding unclassifiable, ranged from 43 to 173 per topic in the integrated data set, with an average of 87.3±48.4, and from 69 to 251 per topic in the separated data set with an average of 124.5±73.9 (Fig. 4). Unclassifiable was the largest category in both the integrated and separated case (337 and 455, respectively) with sex/gender the second-largest (173 and 251, integrated and separated), and religion/cognition the smallest (43 and 59, integrated and separated). The sum of all topic sizes is higher than the number of markings, as markings can be classified in more than one topic.

Variance Analysis

For each topic a p-value was found that represented the degree of clustering. For significance testing, we selected an alpha of 0.05 and have used the Holm's method of correcting for multiple tests (Roback and Askins, 2005). By this method, significance for each potential instance of clustering is as follows: the smallest p-value must be smaller than α/n for that topic to be significantly clustered, the next smallest p-value must be smaller than $\alpha/(n-1)$ for that topic to be significantly clustered, etc. Using these criteria, the integrated data set showed moderate but non-significant clustering in school and sex/gender (p=0.035 and 0.036, respectively, Fig. 5). The variances of all but love/romance frequencies were above the average of the null variance distributions. The estimate of variance in love/romance was in the lower end of the distribution of the null distribution of variances (p=0.68). The separated data set was significantly clustered in five topics: school, sex/gender, religion/cognition, race/national and unclassifiable (p<0.0001, p<0.0001, p<0.0001, p<0.0001, p=0.0033, respectively) (Fig. 6). The estimate of variance in love/romance was just above the median of the null distribution, at p=0.463. Null distributions and p-values are shown in Figure 5 for all specific topics in the integrated data set, in Figure 6 for specific topics in the separated data set, and in Figure 7 for unclassifiable. Figures 5, 6 and 7 show the location of the estimated variance in relation to the null distributions with a bold arrow.

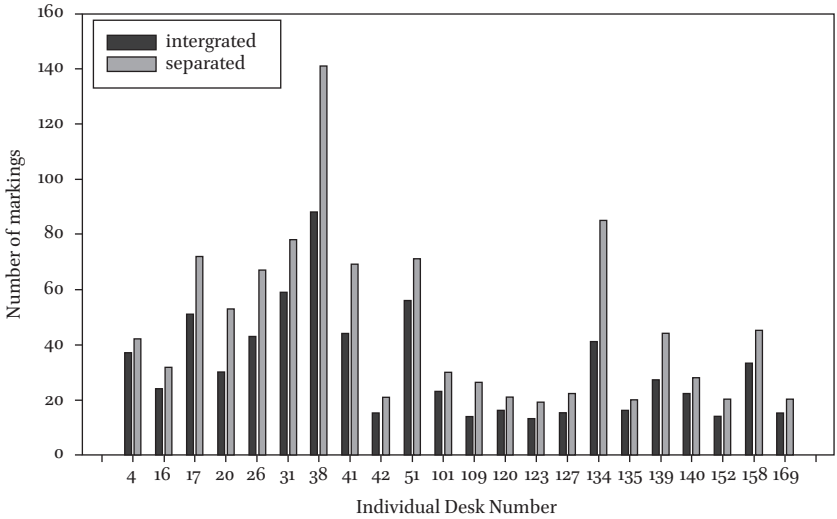

Figure 3. Summary of the number of markings per desk.

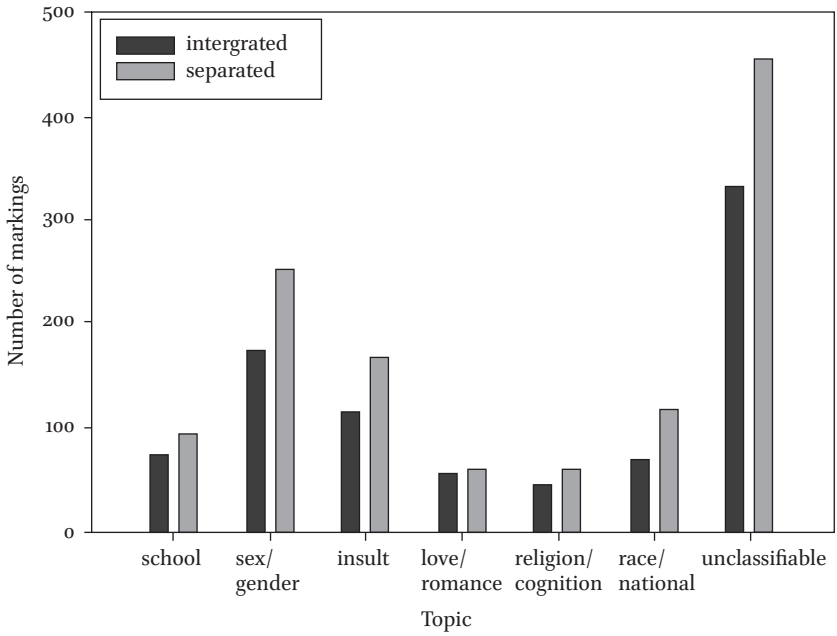

Figure 4. Summary of the number of markings per topic.

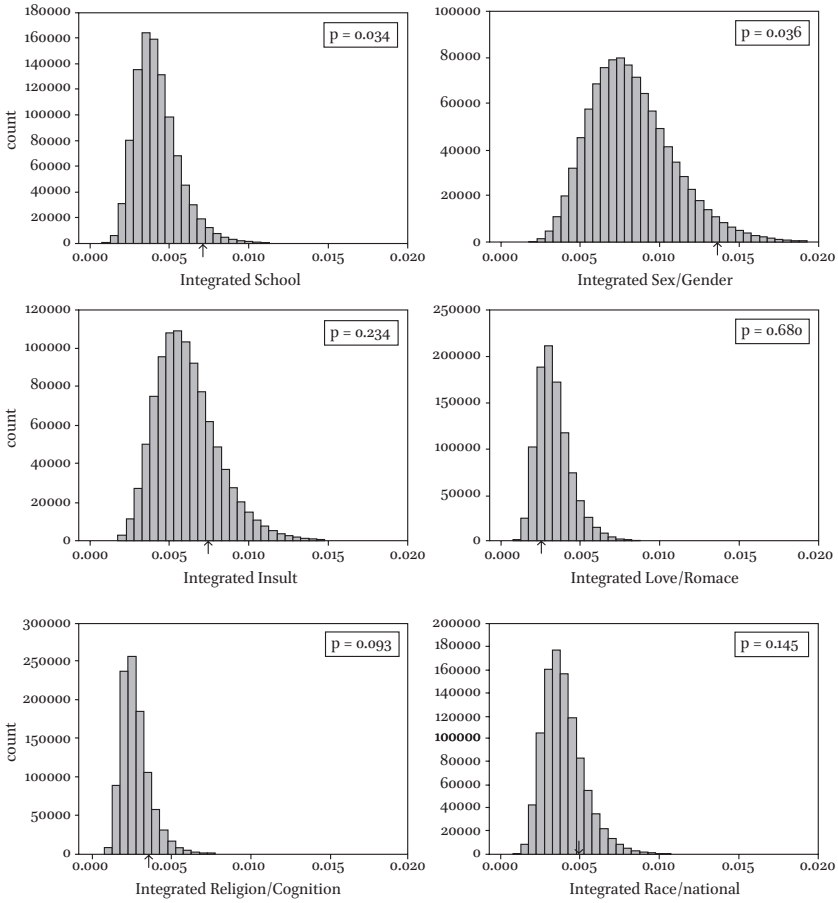

Figure 5. Null distributions of variance values from integrated dataset. Estimated variance is indicated by the bold arrow, *p*-value refers to percentage of distribution above estimated variance.

Topic Correlations

Spearman's ρ correlations between the frequencies of specific topics across desks were almost all non-significant (given an alpha = 0.05 and using the Holm correction for 21 comparisons within the integrated dataset and 21 within the separated dataset), with the exception of insult versus race/national in the separated data set (p=0.0024, Fig. 8). The only other significant correlation was a negative correlation between unclassifiable and insult in the separated dataset. The correlations were generally more strongly negative (but not significantly so) between unclassifiable and other topics (average Spearman's ρ of

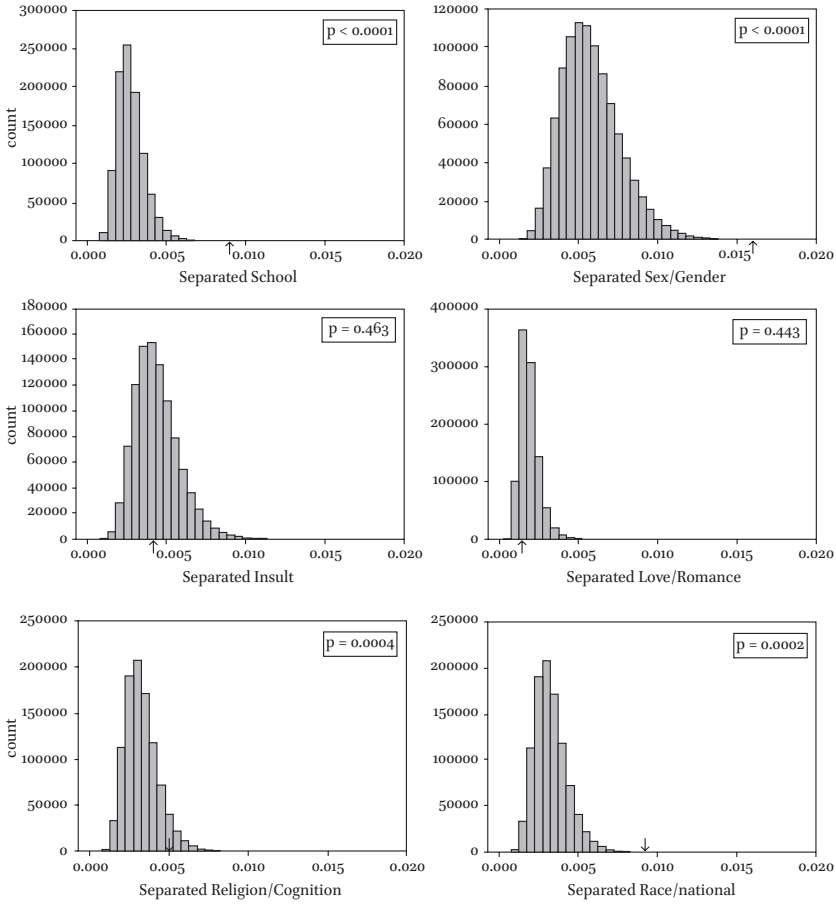

Figure 6. Null distributions of variance values from separated dataset. Estimated variance is indicated by the bold arrow, *p*-value refers to percentage of distribution above estimated variance.

–0.42±0.14 SD for the integrated dataset and –0.37±0.30 SD for separated versus 0.063±0.19 SD between specific topics in the integrated dataset and 0.022±0.25 SD for the separated dataset).

Categorization Survey

The survey was given to three classes with sizes of 20, 30 and 23 for a total of 73 participants. The survey was run as a voluntary exercise, but all people present opted to take it. All participants were undergraduate students at the University of British Columbia, Vancouver, and the majority are likely to be in the faculty

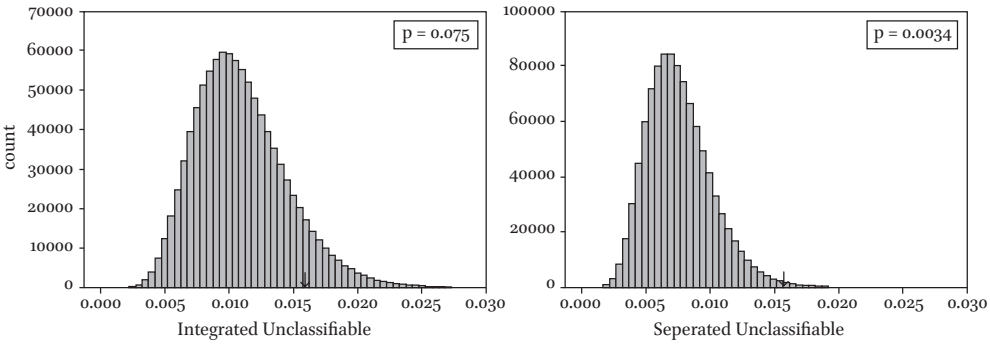

Figure 7. Null distributions of variance values for unclassifiable markings. Estimated variance is indicated by the bold arrow, *p*-value refers to percentage of distribution above estimated variance.

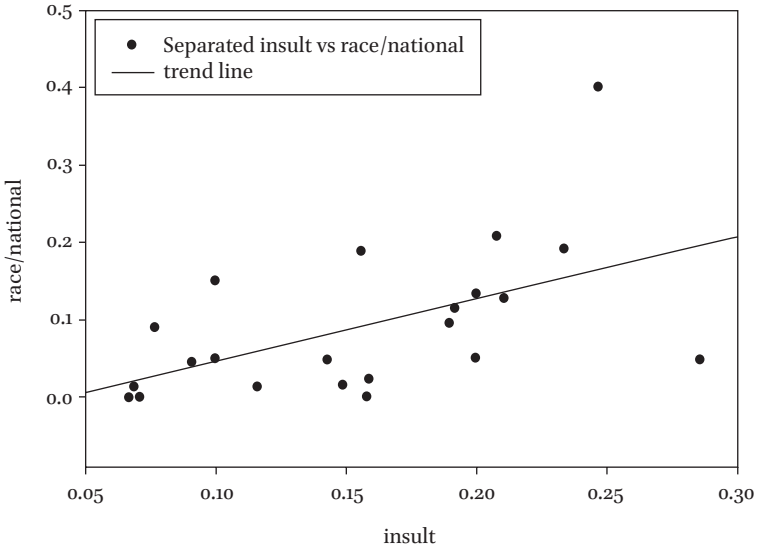

Figure 8. Correlation in frequency between race/national and insult across all desks in the separated (p=0.0024 using Spearman's ρ) dataset.

of science (this was not measured quantitatively, but the classes are all fourth year science courses).

For 29 markings in 7 categories with 3 classes, the total number of classification events (in which a student chooses whether a marking is in a specific category or not) is 606. Out of those 606 classification events, 108 of them were scored positive by either the majority of students or by us. In other words, for 108 of the classification choices, the students or our own method gave a "1" as opposed to a "0". Out of those 108 positive classifications, 64 (59.3% of total positives) were in agreement between the majority of students and us. Out of the 44 remaining positives, 6 (5.6%) were situations in which students placed a marking in a specific topic, whereas we did not. Markings for which we placed a marking in a topic and the majority of students did not were divided into 21 markings that were placed by us into unclassifiable (19.4% of total positives), and 17 markings which were placed by us in a different specific topic than the majority of students. Calculating a true inter-rater reliability score becomes difficult and potentially misleading given the lack of randomization or controlled conditions, so we have left out any further analysis beyond these percentage agreement values.

Discussion

Overview

As an example of cultural dynamics, we have analysed the distribution of phenotypes of graffiti markings that accumulated over a number of years on desks in a university library. Here we view the markings as cultural units in "desk populations" whose dynamics unfold according to certain mechanisms within the background of human psychology. Thus, our perspective is that human psychology constitutes the environment in which cultural units replicate and compete. In our study, the cultural traits or phenotypes of these units are the topics of the graffiti markings, and our principal aim was to test whether topics occurred in clusters, which would suggest either direct or indirect interactions between markings of the same cultural phenotype.

Moderate clustering was observed in 2 out of 6 specific topics (school, sex/gender) in the integrated data set, in which "graffiti conversations" were treated a single unit, and was significant for 4 out of 6 specific topics (school, sex/gender, race/national, religion/cognition) for the separated dataset, in which conversations were split into separate units. The degree of clustering varied with topic, and correlations between specific topics were low, suggesting independence of content between topics. This supports the categorization of

markings into the given topics and the construction of the categories as natural divisions between cultural phenotypes. The following discussion will first address graffiti as a study system, as well as the problem of developing a semantic and objective categorization system for the graffiti markings. We then discuss the finding that some topics appear to have a clustered distribution, as well as the potential mechanisms that can generate such clustering.

Graffiti as a Study System

Graffiti on study desks allows us to have multiple, relatively independent populations of graffiti markings, with each desk representing one population (each desk is small and walled-off from other desks), and each graffiti representing a cultural unit. Graffiti is anonymous, public, and in the case of the library graffiti, can be densely distributed and fairly long-lasting. One of the advantages of anonymity is that it removes most of the effect of prestige-bias, which refers to the preferential copying of specific individuals who are authority figures or otherwise have prestige (Mesoudi et al., 2006; Henrich et al., 2008). Prestige-bias has the potential to complicate cultural unit exchange in the case of live conversation or any instance in which the individuals are identified. In the case of the graffiti we studied, the current distribution of comments is a snapshot that represents an accumulation of comments that has occurred over three years or more, based on information from the library staff (Kevin Madill, Circulation Manager, pers. commun., 2010). This is in contrast to active cultural exchange during a conversation, which is short-lived and more difficult to study, or cultural exchange on the internet, which can be somewhat more complicated to interpret due to search engine dynamics and forum protocols.

Also, graffiti represents cultural exchange on a direct and local scale. To study an evolving distribution of cultural content, a site of active peer-to-peer transmission is an interesting place to focus, and is a more condensed observation of cultural evolution than indirect, long time-scale transmissions, such as those that occur between authors and societies and other authors' books. While both can be considered cultural exchanges, graffiti occurs over shorter time scales and under more direct interactions, and will track cultural evolution in a different way than would following key words and phrases through millions of published texts, as Michel et al. (2011) have done for more than 5 million digitalized books. Graffiti in small study desks provides a good environment in which to study cultural dynamics in an empirical environment across many similar populations of cultural variants.

Library graffiti also has a rich diversity of topics, many of which we have observed to fall into general subcategories. This provides the classifiable cultural variation which is needed to study cultural evolution. While almost no

two markings are the same, the use of general topics as cultural traits is equivalent to focusing on specific phenotypic variables that appear to be relatively general and heritable.

Categorization System

Categorization of semantic categories is a difficult task, because it uses concepts that are necessarily subjective and hard to separate. The content within phrases will frequently fit under multiple categories because of the different ways in which categories can be structured (scale, type, content), and the fact that the same phrase can address multiple ideas. The difficulty in building robust categories is somewhat akin to the difficulty of developing a meaningful definition of species for bacteria, in that there is the problem of horizontal transmission, resulting in networks of genetic history instead of trees (Bapteste et al., 2004). Further, biology has the advantage of being able to focus on the gene, the unit of selection, and watch its movement through populations. For culture, what underlying natural boundaries are there between topics? Aunger (2002) comments that, with the right technology, we may one day be able to define a given topic or cultural trait in terms of the timing and pattern of neural activation in the brain, although this is likely to be different from individual to individual (Mesoudi et al., 2006). We may be a long way from such detailed understanding of a cultural trait. Still, if we accept for now that there is enough of a link between the representations we use for the real world (the representations being the cultural traits) and the differences between physical characteristics in the real world itself, then we can try to develop semantic categories based on this link (i.e., sports versus music as representing different objects and activities in the real world). The categories may overlap, but we believe that this is something we are going to have to accept with cultural traits. The idea of descent from a single parent (from a gene or gene segment's perspective) that is used to develop phylogenetic trees which describe sexual, multicellular, biological descent will not work in many cases for cultural traits.

Graffiti have been categorized in a number of studies, some by the emotions evoked or kind of graffiti (Hagen et al., 1999), others by the content of graffiti (Klofas and Cutshall, 1985; Ball, 2004), or a mixture of both (Şad and Kutlu, 2009; Islam, 2010). Each study validated the selected categories by having multiple people cross-check the categorization and all of these studies have in common the fundamental principle that a good category is one upon which everyone can agree.

Thus, one key concern with our method is that we had only one person develop and define the categories, which may have led to personal bias. The effect of this can be seen in the survey results, which display a slight trend

of overall agreement (more than 50% agreement for all positive classifications), but do not suggest high robustness in categorization (approx. 40% disagreement over positive classifications). Some more ambiguous markings tended to lack any clear agreement between students or between the students and our classification. This could have occurred for different reasons. One source of disagreement is a potential difference in mindset. For the students it may have been less important to them to be consistent in the way they classified each marking, such that they did not attempt to develop internal rules for deciding the specific boundaries of each topic. Our method, on the other hand, followed both a descriptive definition as well as a selection of rules for ambiguous cases (some of which were developed due to information gathered from the survey). Another source of disagreement was differences in cultural background and interpretation of words and context. One example was the marking "You will do great!" which in the context of the library desks, we placed into the category school. The majority of students did not place this marking into school. This may have resulted from different interpretations of the goals and restrictions of topic definitions, something that could potentially have been avoided by giving more explicit and detailed rules to the survey takers. Further, the survey was not a representative sample of the markings available, but a biased selection of the most ambiguous cases. This means that the agreement values can be taken as an underestimate of the percentage agreement that would be expected, had the entire selection of markings been classified by the students.

For our purposes, the way in which an individual marking is placed within a category is connected to the resulting effect on distribution patterns. The categories are semantic categories and as such are built to reflect content, not structure (statement versus question versus doodle, for example). This ties into our prediction that the clustering of certain content is the result of a mechanism in which content is related to fitness, and related to fitness-changes with frequency. A different angle would be to focus on, for example, the length of each marking and consider selection for shorter comments or longer markings (e.g. based on ease of transmission). Thus, it is important to keep in mind that the clustering we observed is a clustering of content, of the specific information communicated by the cultural trait as opposed to the external characteristics or structure of the cultural trait.

Clustering Analysis

Assuming our categories as internally consistent, distinct and natural entities, our results suggest moderate to significant clustering of the topics school and sex/gender. Speculation about why this is the case falls into the realm of psychology. The library is a school library, and these topics may be the most

common thing on students' minds as they sit at desks to study. However, this in itself would not explain the clustering. The idea of school must be somehow inciting, making other graffiti writers think about the topic. Students may already be thinking about school, studying and being in a university library, and there is also an emotional aspect of stress and anxiety. Any school-related comment on the desk would then spark the thought of school easily, and would elicit some understanding or empathy in the reader, whether or not they chose to write.

The emotional state of a student may affect what sorts of topics are strong enough prompts to encourage students to write. Sex/gender may be a similar pattern, in which the emotional (or hormonal) strength behind a topic makes it better at eliciting enough response that graffiti writing occurs. The notion that emotional strength of an idea can determine how likely students are to respond is consistent with the study by Heath et al. (2001), in which the authors show that emotional selection can be stronger than informational selection in selecting which urban legends are passed along. Psychologists from other graffiti studies also suggest that the graffiti reflects the state of mind of people who write it, and can be used to gather information about the general feelings and thoughts of a population who share a space and write graffiti (Islam, 2010).

On the other hand, the topics love/romance and insult showed no tendency for clustering, suggesting that there may be something different about the psychological environment of these topics, or of the functional quality of their categorization. By this we mean that love/romance may be a topic that does not spark response, but is fairly frequently on the minds of students. People may write "JH <3 KT" but not necessarily feel the need to respond the same way they might be emotionally incited to respond to a school or sex/gender comment. This may have something to do with the personal nature of romantic relationships, as opposed to the generality of school and sex, but this is pure speculation. Again, this may both inform us about love/romance as a cultural trait, and reveal potential pressures due to the cultural environment of human psychology.

For the topic insult, something similar to love/romance may have been occurring, in that insults were personal attacks as opposed to specific discussion, and so could be in response to anything. Thus, many insults would be no more likely to spark another than a single insult. This lack of clustering may also result from the kind of category that *insult* is, because it describes a form of comment or communication, and is less of a content-based category. To further explore if this specific effect, of having functional or descriptive categories lack clustering where content-based categories show clustering, it would be useful

to expand the number of different categories. Such a study would inform the issue of category definition as well as shed light on whether it is the content that is replicating or the form of category.

Separating conversations had a substantial effect on the degree of clustering. The overall increase in clustering with conversation separation suggests that the topics that occurred within conversations tended to be related, and increased the level of clustering in general. This makes sense intuitively, as conversations are direct responses to a given marking, as opposed to simply additions to the desk that appear to be intended as independent contributions. It would seem less likely for someone to respond to a conversation about exam grades with a non-sequitur about sports or religion, though topic changes within conversations did occur to some extent. The same bias may occur within conversations, but be stronger than for markings added in isolation.

Topic Correlation

The only significant topic correlations were race/national with insult and unclassifiable with insult in the separated dataset. The lack of correlation between most specific topics and a trend towards negative correlation with unclassifiable gives support to the independence of each category and suggests the robustness of each category as a distinct cultural unit. The correlation between race/national and insult is reasonable even in the context of distinct categories. There were many markings that were either entirely about race but were not insulting, and insults that had nothing to do with race, supporting the notion of distinct categories. Still, many of the race/national markings were insulting, and some comments which were racial but not directly insulting were responded to with reactionary insults.

Mechanisms for Clustering

There are several potential mechanisms for clustering. For example, "conformity bias" refers to behaviour in which individuals adopt a specific cognitive attractor based on which attractor is in the majority in the surrounding population. Conformity bias is considered a mechanism of cultural positive frequency dependence (Mesoudi and Lycett, 2008) and can counter the weakening of selection through inaccuracy of transmission by decreasing variation in a way similar to selection (Henrich and Boyd, 2002). Various studies with specific focuses in archaeology and artefact lineages use the idea of conformity bias to explain patterns of cultural descent (Eerkens and Lipo, 2005; Charlton et al., 2010; Shennan, 2011). Garland et al. (2011) recently reported that male humpback whales show strong conformity bias in the song they choose to sing for a given

year. Each year or two, a single song will sweep from west to east across the globe (taking a few years to make the journey), being either completely unique from the song of previous years or a slight modification on similar themes (Garland et al., 2011). In addition to conformity bias, there may be another content-based mechanism at play which gives a proportionate advantage to topics in the majority, and this is semantic priming.

Semantic priming is a non-biased, unconscious mechanism which also has the potential to lead to clustering in situations of spatially and/or temporally discrete cultural exchange. Priming in general is a psychological effect in which previous stimuli to an individual affects the response to future stimuli (Tulving and Schacter, 1990; Ferrand and New, 2004). Semantic priming specifically refers to priming that occurs based on meaning, as opposed to sound or historical context of a phrase. For example, priming with the word "wolf" can result in participants being more likely to answer the question "what starts with D?" with the word "dog", because the terms are semantically related. The neurological mechanism of priming is under some debate, but it is agreed that priming is unconscious and mostly independent of other kinds of memory (Tulving and Schacter, 1990; Levy et al., 2004; Squire, 2004). Disentangling conformity bias and semantic priming as causes for the clustering observed in our study is beyond the scope of this work.

In addition, there are other possible mechanisms for clustering beyond conformity bias and semantic priming. These could include similar students using the same desks and tending to have the same topics on their minds. Different handwriting suggests that this isn't generally the case, but it cannot be ruled out completely. It is possible, too, that something about the location of specific desks had an effect on student psychology. For example, desks near the back of the library may either attract students of a certain personality type or induce different attitudes than desks at the front. To examine spatial effects, it would be interesting to look for spatial clustering in topics at a scale larger than individual desks, i.e., clusters of desks or regions of the library. Because desks are fairly separated from each other physically, strong clustering at a broader scale would suggest a mechanism different from priming or conformity bias, suggesting instead a pattern resulting from psychological cues of an area or similar students returning to the same sets of desks. However, the desk organization is complex enough that a proper spatial analysis is outside the realm of this study. Calculating the distance between desks in order to measure patterns in topic distribution is made difficult by the numerous obstacles and pathways between desks (i.e., it is not a simple grid structure). A qualitative map of topic frequencies is provided in Appendix A to show the complexity of the desk arrangements and the potential patterns that may exist. It would also be highly informative to conduct an experiment which manipulates topic frequencies on desks and

records changes over time. This would reveal the curve of topic population growth over time, as well as the strength of selection for topics of different relative frequencies on the same desk, both of which would increase our understanding of the mechanisms causing the observed clustering.

In addition, focusing on categories of lower frequency overall may reveal trends of negative frequency dependence, which in the case of cultural traits is essentially non-conformity bias. Negative frequency dependence has been modelled with respect to the diversification of religions by Doebeli and Ispolatov (2010), and the distribution of cultural variants under negative frequency dependence was also modelled by Mesoudi and Lycett (2008). However, to our knowledge there is a limited amount of empirical work on negative frequency dependence in cultural traits, and graffiti on study desks has the potential to be a useful study system for this subject as well.

Conclusion

The distribution of topics of graffiti marking on library desks showed clustering for some topics, but not for others. This suggests that certain cultural traits have some kind of direct or indirect relationship between "individuals" within the same trait or topic in the graffiti environment. It would be interesting to test this with experimental studies that explicitly capture the dynamics of graffiti populations over time. Such studies could for example involve the priming of desks with graffiti of certain content categories. Thus, our study could serve as the basis for further investigations of the dynamics of cultural evolution in a system that accumulates cultural traits over space and time. Categorization of cultural traits is a difficult process. It is clear that more detailed methods accommodating overlap and nesting between topics and blurry boundaries will be an asset to the study of cultural evolution. Our study shows that it is possible to apply concepts from biology (such as the analysis of the dynamics of phenotype frequencies) to cultural evolution, once we have defined traits or phenotypes on which to base such analyses. Defining those traits in a meaningful way is key to making any analysis informative of the true dynamics of cultural systems.

Acknowledgements

We thank Michael Whitlock for providing statistical advice, and Erik Hanschen for discussions. M.D. was supported by NSERC (Canada) and by the Human Frontier Science Program.

References

Aunger, R. A. (2002). *The Electric Meme: A New Theory of How We Think*. Simon and Schuster/Free Press, New York, NY.

Ball, D. B. (2004). *The Nature and Extent of Desktop Graffiti Among U.S. College Students: An Exploratory Study*. Masters Thesis. Virginia Tech University, Blacksburg, VA.

Bapteste, E., Boucher, Y., Leigh, J. and Doolittle, W. F. (2004). Phylogenetic reconstruction and lateral gene transfer. *Trends in Microbiology* 12, 406-411.

Boyd, R., Richerson, P. J. and Henrich, J. (2011). Rapid cultural adaptation can facilitate the evolution of large- scale cooperation. *Behavioural Ecology and Sociobiology* 65, 431-444.

Cardoso, G. C. and Atwell, J. W. (2010). Directional cultural change by modification and replacement of memes. *Evolution* 65, 295–300.

Charlton, M. F., Crew, P., Rehren, T. and Shennan, S. (2010). Explaining the evolution of ironmaking recipes – an example from the northwest Wales. *Journal of Anthropological Archaeology* 29, 352-367.

Doebeli, M. and Ispolatov, I. (2010). A model for the evolutionary diversification of religions. *Journal of Theoretical Biology* 267, 676-684.

Eerkens, J. W. and Lipo, C. P. (2005). Cultural transmission, copying errors, and the generation of variation in material culture and the archaeological record. *Journal of Anthropological Archaeology* 24, 316-334.

Ferrand, L. and Boris, N. (2004). Semantic and associative priming in the mental lexicon. In Bonin, L. (Ed.), *The Mental Lexicon*, pp. 25-43. Nova Science Publishers New York, NY.

Friesen, M. and Bailey, D. (2010). UBC Library Services Statistics 2009–10. UBC Library. Available online at https://circle.ubc.ca/handle/2429/26153 (accessed 2 April 2011).

Garland, E. C., Goldizen, A. W., Rekdahl, M. L., Constantine, R., Garrigue, C., Hauser, N. D., Poole, M. M., Robbins, J. and Noad, M. J. (2011). Dynamic horizontal cultural transmission of humpback whale song at the ocean basin scale. *Current Biology* 21, 1-5.

Hagen, C. A., Ender, M. G., Tiemann, K. A. and Hagen, C. O. (1999). Graffiti on the Great Plains: A social reaction to the Red River Valley flood of 1997. *Applied Behavioral Science Review* 7, 145-158.

Heath, C., Bell, C. and Sternberg, E. (2001). Emotional selection in memes: the case of urban legends. *Journal of Personality and Social Psychology* 81, 1028-1041.

Henrich, J. and Boyd, R. (2002). On modeling cultural evolution: why replicators are not necessary for cultural evolution. *Journal of Cognition and Culture* 2, 87-112.

—— and Richerson, P. J. (2008). Five misunderstandings about cultural evolution. *Human Nature* 19, 119-137.

Horner, V., Whiten, A., Flynn, E. and de Waal, F. B. M. (2006). Faithful replication of foraging techniques along cultural transmission chains by chimpanzees and children. *Proceedings of the National Academy of Sciences of the United States of America* 103, 13878-13883.

Islam, G. (2010). Backstage discourse and the emergence of organizational voices: exploring graffiti and organization. *Journal of Management Inquiry* 19, 246-260.

Klofas, J. and Cutshall, C. (1985). Unobtrusive Research methods in criminal justice: using graffiti in the reconstruction of institutional cultures. *Journal of Research in Crime and Delinquency* 22, 355-373.

Levy, D. A., Stark, C. E. L. and Squire, L. R. (2004). Intact conceptual priming in the absence of declarative memory. *Psychological Science* 15, 680-686.

Mesoudi, A. and Lycett, S. J. (2009). Random copying, frequency-dependent copying and culture change. *Evolution and Human Behavior* 30, 41-48.

Mesoudi, A. and Whiten, A. (2008). Establishing an experimental science of culture: animal social diffusion experiments. *Philosophical Transactions of the Royal Society B* 363, 3477-3488.

Mesoudi, A., Whiten, A. and Laland, K. N. (2006). Towards a unified science of cultural evolution. *Behavioral and Brain Sciences* 29, 329-383.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A. and Aiden, L. E. (2011). Quantitative analysis of culture using millions of digitized books. *Science* 331, 176-182.

OpenOffice.org. (2010). Available online at http://www.openoffice.org/.

Roback, P. J. and Askins, R. A. (2005). Judicious use of multiple hypothesis tests. *Conservation Biology* 19, 261-267.

Rogers, D. S. and Ehrlich, P. R. (2008). Natural selection and cultural rates of change. *Proceedings of the National Academy of Sciences of the United States of America* 105, 3416-3420.

Şad, S. N. and Kutlu, M. (2009). A study of graffiti in teacher education. *Eurasian Journal of Educational Research* 36, 39-56.

SAS Institute (1989-2009). *JMP*, Version 8. SAS Institute, Cary, NC.

Shennan, S. (2011). Descent with modification and the archaeological record. *Philosophical Transactions of the Royal Society B* 366, 1070-1079.

Squire, L. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory* 82, 171-177.

Tulving, E. and Schacter, D. L. (1990). Priming and human memory systems. *Science* 247, 301-306.

Appendix A

Floor map of Koerner library basement study area. Darker grey marks desk areas, lighter grey marks bookshelves, walls, and other obstacles. Pie charts are located near the desks they represent, with topic legend in upper left. Population size of each desk is not represented in this image. This figure is published in colour in the online edition of this journal, which can be accessed via http://booksandjournals.brillonline.com/content/15685373.