Supporting Information

Hughes et al. 10.1073/pnas.1115407109

SI Text

List of Function Words Used. In our experiments, we used the 307 function words listed in Table S1 to measure style in the works considered.

Style Network. In our experiments, we found preliminary evidence that the strong (i.e., statistically significant) stylistic connections between authors, although generally reflective of a "style of a time," also show grouping based on thematic similarity. Thematic connections can provide a substrate for the transmission or reification of style, through selective samplings of text based on themes. The tendency of authors to cluster in thematic ways in shown in Fig. S1, which displays a network representation of the stylistic connections between authors that were statistically significant at the $\alpha = 0.002$ level. (These connections were chosen just as before using the pairwise similarities derived from KL divergence between the function word frequency author feature vectors.) Within this network, we have magnified several groupings of authors that reflect thematic or genre-based similarities, in particular the group of English poets and playwrights that includes Christopher Marlowe and William Shakespeare, a separate (i.e., disconnected) component of Civil War generals, and the group of naturalists, philosophers, and social thinkers that includes Charles Darwin, Thomas Huxley, and Bertrand Russell, among others.

Robustness Analysis. We consider the robustness of our results over the period 1836–present, using the year at which we began to see a superlinear increase in the number of authors per unit increase in year as a starting point. We performed three analyses of the trends in similarity over this period, one using all of the authors therein, one that subsampled the densest period (1836–1924) by including only every fifth author, and one that subsampled the densest period by including every eighth author. The last of these effects a normalization of the density of authors that gives us a density that is roughly equal across the entire time span of our dataset. Note that while we considered temporal windows only within the period 1836—present, for our average similarity calculations, we included all previous authors, including those that fell outside that period.

Fig. S2 shows the result of this robustness analysis. Therein we see the (subsampled) "influence surfaces" as well as the "original" influence surface—i.e., a figure produced using the same procedure but without any subsampling. Note that these are smoothed versions of the average similarity surface, where the average was computed at each author year (and temporal window size) by bootstrap sampling of all author similarities that fell within that window using 100 runs and 1,000 samples per run. A time slice of any of these surfaces (i.e., the graph obtained for a fixed author year) shows the influence trend (i.e., similarity score moving back in time) for an author of that particular year. Note that the three subfigures have the same general shape, consistent with claims of robustness (to sampling and density variation).

A Simple Evolutionary Model for Influence. Our findings raise the interesting question of what is the mechanism for the observed decline in similarity (influence). One simple model is as follows. Assume that the number of books that any one individual can read in a life time is a constant, K. Over time the number of books that are published has been growing exponentially at a rate e^{rt}, where the rate parameter r reflects a positive increase in rate of publishing. Assuming that all books remain in press, at any time t the number books in circulation will be B_0e^{rt}, where B_0 is the number of books available at the start of book publishing. We define an interval of time $\delta t = t(n-1)$, where the variable $n \geq 1$ is the multiple of time in years over which an individual can read without exceeding her book capacity K. Hence the value of n must be such that the following equality holds:

$$K = B_0(e^{rnt} - e^{rt}).$$

This implies that the numbers of years back from a present moment over which all books can be read is given by,

$$n = 1 + \frac{1}{rt}\log\left(\frac{K}{B_0}\right).$$

It is evident that for any positive rate of book growth, $r > 0$, as $t \to \infty$, $n \to 1$. Hence as we move into the future the interval of time over which we can read will diminish, $\delta t \to 0$.

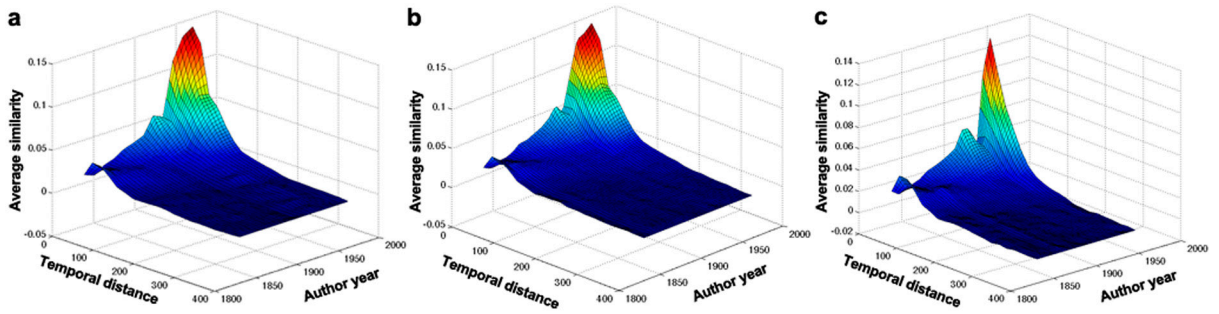

Fig. S1. Subsampled influence surfaces. These surfaces show the full family of influence curves obtained from subsampling the Gutenberg data by including every eighth author (*A*), every fifth author (*B*), and every author (*C*). Note the same general shape of the surfaces, which are each consistent with the influence curves from the full data.

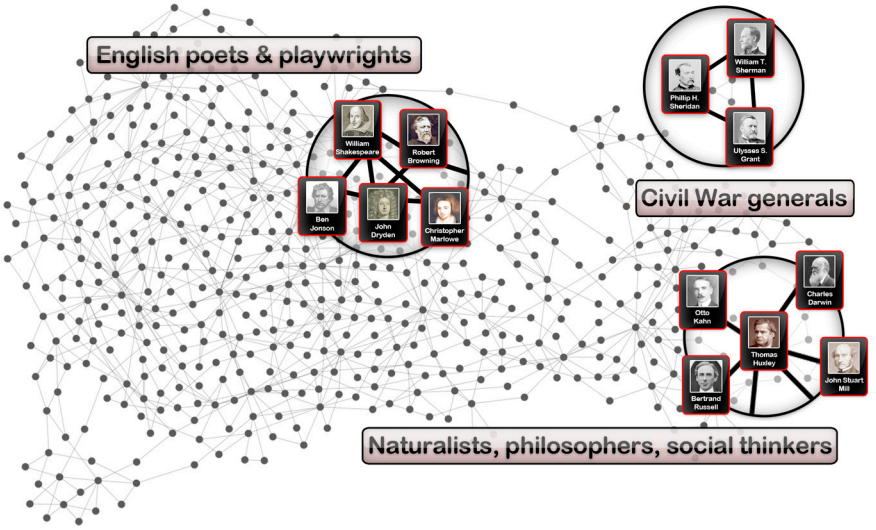

Fig. S2. Network representation of statistically significantly similar styles among the 537 authors considered at the $\alpha = 0.002$ level. Shown are three clusters of authors with significant connections that also reflect thematic clustering, in some cases superseding the temporal distance between the authors.

Table S1. List of 307 function words used to measure style in the works considered

a	about	above	across	after	afterwards	again
against	all	almost	alone	along	already	also
although	always	am	among	amongst	amoungst	amount
an	and	another	any	anyhow	anyone	anything
anyway	anywhere	are	around	as	at	back
be	became	because	become	becomes	becoming	been
before	beforehand	behind	being	below	beside	besides
between	beyond	both	bottom	but	by	call
can	cannot	cant	con	could	couldnt	cry
describe	detail	do	done	down	due	during
each	eight	either	eleven	else	elsewhere	empty
enough	etc	even	ever	every	everyone	everything
everywhere	except	few	fifteen	fify	fill	find
fire	first	five	for	former	formerly	forty
found	four	from	front	full	further	get
give	go	had	has	hasnt	have	he
hence	her	here	hereafter	hereby	herein	hereupon
hers	herself	him	himself	his	how	however
hundred	ie	if	in	inc	indeed	into
is	it	its	itself	keep	last	latter
latterly	least	less	ltd	made	many	may
me	meanwhile	might	mine	more	moreover	most
mostly	move	much	must	my	myself	name
namely	neither	never	nevertheless	next	nine	no
nobody	none	noone	nor	not	nothing	now
nowhere	of	off	often	on	once	one
only	onto	or	other	others	otherwise	our
ours	ourselves	out	over	own	part	per
perhaps	please	put	rather	re	same	see
seem	seemed	seeming	seems	serious	several	she
should	show	side	since	six	sixty	so
some	somehow	someone	something	sometime	sometimes	somewhere
still	such	take	ten	than	that	the
their	them	themselves	then	thence	there	thereafter
thereby	therefore	therein	thereupon	these	they	thin
third	this	those	though	three	through	throughout
thru	thus	to	together	too	top	toward
towards	twelve	twenty	two	under	until	up
upon	us	very	via	was	we	well
were	what	whatever	when	whence	whenever	where
whereafter	whereas	whereby	wherein	whereupon	wherever	whether
which	while	whither	who	whoever	whole	whom
whose	why	will	with	within	without	would
yet	you	your	yours	yourself	yourselves	