Current Comments

# The Epidemiology of Knowledge and the Spread of Scientific Information

Number 35 September 1, 1980

People commonly refer to "infectious" slogans, "catchy" phrases, or ideas that "spread like wildfire." These everyday expressions indicate a basic similarity between the dissemination of ideas and the transmission of disease—both are natural processes in which something is communicated, that is, transferred from one person to another. When a disease spreads quickly and infects many individuals it is called an epidemic. Ideas, too, can spread quickly and "infect" many people. I had the opportunity to explore the analogy between medical and "intellectual" epidemics in a lecture I gave last December at the Center for Disease Control in Atlanta, Georgia.[1]

I first heard about this analogy from my old friend Bill Goffman, even before he published the primordial paper in *Nature*, in 1964.[2] In that paper, Goffman and Vaun Newill, Case Western Reserve University, Cleveland, Ohio, pointed out that the dissemination of scientific ideas could usefully be described as a process similar to the transmission of disease. In fact, they suggested that existing mathematical models which describe epidemic processes could be valuable tools for information scientists as well as for medical researchers. Medical researchers use epidemic models both to *describe* the spread of a disease within a population and to *predict* when it is likely to reach a peak of infection, after which it presumably will decline. Goffman and Newill proposed that the same models could also describe the spread of information within the research community and predict the probable duration and intensity of intellectual epidemics. Working alone or with various colleagues, Goffman has tested this proposal by applying the epidemic model to the literature of different fields.

To draw his analogy between medical and literature epidemics, Goffman identifies the elements of all epidemics. The first element is the infectious material itself, and how it is communicated. In medical epidemics, the infectious material is a virus, bacterium, parasite, fungus, or whatever. Exposure to these organisms is either direct or indirect. For example, venereal disease is transmitted by direct physical contact while malaria requires an intermediate "vector"—the mosquito—for transmission to humans. In intellectual epidemics, *ideas* are the infectious material. Ideas are communicated informally through conversations or seminars, for example, or ideas may spread through a journal "vector" by formal publication.[2]

The second element of epidemics is the population through which they spread. Members of the population belong to one of three categories at any specified time: infectives, susceptibles, and removals. Infectives are those who harbor the infectious material. In medical epidemics, infectives are people who carry the bacteria, or whatever. In intellectual epidemics, authors or researchers are infectives who have ideas to communicate. Susceptibles are those who come in contact with the infectious material—anyone who may come down with the infectious disease, and journal readers or conference attendees who encounter the idea. Removals are former infectives or susceptibles—those who are immune to the disease, are hospitalized for treatment, or have died from it, and those who resist ideas or are no longer active researchers because of retirement or death.

The epidemic process itself may be stable or unstable. In a stable state, the number of infectives doesn't increase or decrease over time. When the disease process is stable, it is considered to be in an "endemic" state. In an unstable state, the number of infectives may be increasing, in which case the epidemic is spreading, or it may be decreasing as the epidemic declines over time.[3]

A complex series of differential equations is applied to calculate the rates of change in the number of infectives, removals, and susceptibles.[4] I won't take time to describe them here because they aren't essential to this essay. However, it is important to note that these equations solve three problems. First, they define the "curve" which traces the growth of the epidemic over time. Second, they define the conditions under which the epidemic reaches a peak point and stabilizes. Third, they define the threshold number of susceptibles which must be exceeded if the epidemic is to begin.[3]

Of course, all mathematical models are simplifications of rather complicated real-world problems.[5] Goffman points out that an "almost endless number of complexities" emerge which the model for epidemics, both medical and intellectual, cannot take into account.[2] Although many details may be passed over, the essential features of the actual process are described by the mathematical model. Goffman has applied the epidemic model to the literature of several fields to see how well it accounts for the nature of scientific growth and the spread of information. In particular, Goffman hoped to determine, quantitatively, the relative importance of different areas within a field and to predict the future direction of research in the field.

Goffman's first literature study concentrated on the spread of knowledge about mast cells.[6] He defined the basic population as the total number of authors listed in a bibliography compiled by Hans Selye. Selye's bibliography included "all the contributions to the subject area, from Ehrlich's discovery of the mast cell in 1897 until 1963."[6] The bibliography lists 2,195 authors and 2,282 publications. Goffman considered the growth of this literature as a "two-factor" epidemic process involving the direct transmission of ideas between authors without consideration of the journal "vector." The authors were classified as infectives or removals. Authors became infectives in the first year of publication of their papers on some specific aspect of mast cell research. Authors became removals one year after the date of publication of their last paper in the Selye bibliography.

Goffman plotted the rates of change over time of both the number of authors and the number of publications. The curves indicated that changes in the number of publications mirrored those for authors, and the ratio of publications to authors was constant. Thus, the epidemic "literature explosion" of mast cell research is at the same time a "population explosion" of authors—the spread of infectious materials (papers) is proportional to the increase in infectives (authors).[6]

The curves also showed that mast cell research was fairly stable for almost 60 years after Ehrlich's initial "infective" discovery. The epidemic model requires that the rate of change of removals be constant when the process is stable. Goffman plotted this rate and found that it was indeed stable, showing that the epidemic model is suitable for analyzing the spread of mast cell research.

After the 60 year period, a sharp and steady increase in both total authors and publications was evident. Thus, mast cell research achieved epidemic proportions around 1940 after a 60 year "latency period" of stability.[6]

When Goffman analyzed the epidemic phase of mast cell research, he found it actually constituted three separate lines of investigation. The first originated in the discovery by Scandinavian workers that mast cells of certain species stored and synthesized heparin. The second spread from research centered in Scotland demonstrating the histamine content of mast cells. The last developed from American research showing that these cells contain serotonin. Goffman treated each line of investigation as a separate epidemic process and concluded that the histamine epidemic was the most "virulent" in terms of size and intensity.[6]

Goffman later applied the epidemic model to analyze the development of symbolic logic from 1847 to 1962.[7,8] The literature population for this study was compiled from a comprehensive bibliography covering the period 1847-1935, as well as journal material from 1936-62. In all, the population consisted of 1,733 authors and 5,845 publications. As in the mast cell study, authors were defined as infectives in the first year of their publication in the field or as removals one year after their last paper appeared in the population.

When the changes in total authors and total publications were plotted over time, the resulting curves again showed that the ratio of papers to authors was constant. Also, symbolic logic entered an epidemic phase in 1892, 45 years after the germinal publications of Boole and DeMorgan in 1847. The spread of symbolic logic research actually was a *recurring* epidemic process, with peaks occurring every 25 years.

Goffman traced the sources of each recurring epidemic to smaller specialties within the field of symbolic logic. He found that the reasons for their peaks and declines conformed to epidemic theory. For example, metamathematics was one of five specialties triggering an epidemic of research activity in 1922 that peaked in 1932. Although the curve indicated that metamathematics was not in an epidemic phase in 1932, the number of susceptibles at that time exceeded the threshold set by epidemic theory. Thus, Goffman predicted that metamathematics would enter an epidemic phase in 1932, and it did.

Both the symbolic logic and mast cell literature epidemics showed that the ratio of papers to authors was constant. Thus, the number of authors contributing to research in these fields rapidly increased during the epidemic phase and quickly dropped off when the epidemic waned. Obviously, some proportion of these infective authors contribute only one or two papers before they become removals.

The extent of such author "turnover" was quantitatively measured by Donald Hawkins of Bell Laboratories in Murray Hill, New Jersey, in his examination of the growth of literature on noble gas compounds.[9] The literature population consisted of 1,123 authors and 1,192 references published from 1962 to 1977. Of the 1,123 authors, 703 (66.4%) were active for only one year! Only 53 authors (5%) were active for ten years or more. Hawkins concluded that "the unusual features of the noble gas compound literature are its sudden start, rapid growth, and great interest to a large number of investigators. Most of these chemists remained active in this field for only a short time."[9]

However, high author turnover in scientific research may be the rule, not the exception. Derek Price and Suha Gürsey, Yale University, randomly selected 506 authors from the *Science Citation Index® (SCI®)* from the period 1964-70.[10,11] This represented a sample population of "a little more than a million scientific authors in all the countries of the world."[10] Of the 506 authors listed, 281 names (56%) occurred in only one of the *SCI* source author indexes—that is, 56% of the authors listed were "transients" who were active for only one year. Nineteen authors (4%) were active for all seven years, representing a small core group of "continuant" authors. Significantly, the small group of continuants "will probably produce more than half the total output [of research publications]."[10] Price and Gürsey believe that the large number of transient researchers is an important phenomenon for sociologists of science to explore.

As I said earlier, Goffman treated the epidemic growth of the symbolic logic and mast cell literatures as a two-factor process without considering the role of the journal, which acts as a "vector" carrying published information between infectives and susceptibles. Goffman originally proposed a three-factor epidemic model that included the journal vector,[2] but it wasn't actually tested against a body of scientific literature. However, working with Kenneth Warren, who was then affiliated with Case Western Reserve and is now at the Rockefeller Foundation, Goffman suggested that a *four*-factor model would best describe the exponential growth of medical literature.[12] Recently, Goffman and Warren published a book detailing the four-factor model and reviewing past epidemiological literature studies.[13]

The four-factor model for medical literature growth emerged from an earlier article by Goffman and Warren on the epidemiology of schistosomiasis.[14] In schistosomiasis, there are two hosts, definitive and intermediary. Man is the definitive host and the intermediary is a specific species of snail. The infectious agent is a parasite. In one stage of its life cycle (cercaria), it is infective for man and in another (miracidium), it infects snails. The disease process is a cyclic phenomenon: the parasitic worm in man lays eggs which leave the body with his excretion; the eggs mature and hatch in fresh water, where they penetrate the snail; after transforming into cercariae, the parasites leave the snail, return to the water where they penetrate the skin of man, and again lay eggs in the blood vessels of the intestines and urinary bladder.[14]

Health officials have an obvious interest in limiting schistosomiasis, which annually affects 200 million people worldwide.[12] The mathematical equations that solve the four-factor epidemic process give health officials important clues as to how schistosomiasis may be controlled. Goffman and Warren point out that "there exists a threshold above which the intermediate population ...must pass for an epidemic outbreak to occur; and there is a linear relationship between infectives in the definitive and intermediate host populations. It is obvious that the infectious process can be controlled by controlling the intermediate host population."[12] But they also caution that the cyclic interaction makes schistosomiasis an *ecological* process—various biological factors coexist in a balanced environment, and changes in any single factor will upset the whole, for better or worse.[12]

The spread of medical information is also a cyclic process involving four factors. The definitive host is the researcher and the journal is the intermediate host. The infectious material, information, has two stages of development—the manuscript and the finished article. The researcher develops an idea which is released in the academic community in the form of a manuscript; the manuscript is accepted by a journal where it is edited and transformed into a finished paper; other susceptible researchers come in contact with the idea, and they become removals if they reject the idea or infectives if they accept the idea and cite the paper in their own research.[12]

The authors demonstrated that the spread of information in the medical community is also an ecological process in which the various separate factors are all related. They analyzed the literature on schistosomiasis from 1852 to 1962, totalling more than 10,000 publications. They also asked 47 experts in the field to qualitatively review the bibliography. The experts identified more than 3,000 "quality" publications, each of which was chosen between one and 25 times. The quantitative and qualitative analyses of the schistosomiasis literature yielded some interesting results.[12]

Like many other fields, schistosomiasis publications are still growing exponentially. As in the earlier studies on symbolic logic and mast cells, the paper-to-author ratio is constant over time. Also, the number of authors determines the number of journals publishing the schistosomiasis literature. Finally, the number of authors producing "quality" papers is directly proportional to the total number of authors in the field. Thus, in the ecology of medical literatures, the numbers of authors, journals, articles, and "quality" articles are related—a change in one will affect the balance of the whole.[12]

The exponential growth of scientific literature raises questions about whether the journal publication system should be changed or, at least, augmented by alternative channels of communication. Four-factor epidemic theory indicates that the spread and volume of infectious information can be controlled simply by limiting the intermediate host (journal) population. However, Goffman and Warren warn that the result may be a decline in both the total population of infective authors and the proportion of quality research.[12]

Instead of tampering with the ecological system of journal publication, they conclude it would be less disruptive if susceptible researchers control the amount and quality of infectious material with which they come in contact. Unlike disease processes which must be limited for obvious reasons, the dissemination of information should be encouraged to spread to all members of a population who may find it useful. The problem is not how to limit the exponential growth of scientific literature, but how to ensure that the most relevant and most virulent information is communicated.[12] ISI® and other information retrieval services are designed for this very purpose, and they may be used in conjunction with Goffman's and Warren's suggestions. However, our most-cited article studies are particularly designed to foster this process. We are also interested in learning why some important ideas do not become virulent for many years.

Contrary to much of the accepted wisdom, Goffman and Warren advise susceptible researchers to reduce the amount of primary literature they regularly scan in order to avoid acquiring an "immunity of incomprehension" from constant overexposure.[12] They recommend that researchers first establish a core list of journals for their major areas of interest. They observe that Bradford's law of literature scattering, which I've discussed many times,[15] supports the idea that a relatively small nucleus of journals accounts for a large proportion of a field's literature. The *Journal Citation Reports*®[16] (*JCR*™) volume of the *SCI* can help identify the core journals once you specify one or more particular journals in your field.

The literature to which a researcher is exposed should also be "of a reasonably high degree of virulence."[12] This is just another way of saying you should read the high impact papers. Goffman and Warren suggest the organization of "quality review" panels within specialist scientific societies. These panels would alert members of the society to significant articles appearing in journals they may not regularly read. In a way, this is how *Excerpta Medica* (*EM*) is organized. *EM* relies on biomedical and clinical specialists to canvass the world scientific literature for articles to appear in its abstract journals covering 43 separate specialities.[17] At ISI, our citation studies identify particularly virulent articles and highly infective individuals in terms of the number of citations each accrues over time.[18,19]

Also, ISI's co-citation analysis and cluster mapping procedures represent another approach to studying the epidemiology of scientific literature.[20] By looking at cluster maps generated on an annual schedule, the literature epidemiologist could see how an outbreak of research actually spreads over time, through what channels, and to whom. Perhaps the mathematical models proposed by Goffman and others can be combined and applied to ISI's analysis and mapping to predict when an area of research is likely to break out in epidemic proportions, how long the epidemic may last, how many people may be infected, and when an information retrieval system should be introduced to facilitate the communication of relevant scientific information.[21] ISI's and other data bases may serve as a "real world" test for the various mathematical models of intellectual epidemiology. Just as we hope one day to control diseases better through such models, we may be able to optimize information flow by the kind of thinking Bill Goffman has pioneered.

* * * * *

*My thanks to Patricia Heller and Alfred Welljams-Dorof for their help in the preparation of this essay.*

## REFERENCES

1. **Garfield E.** *Evaluating information collecting systems.* Unpublished lecture presented at the Center for Disease Control. 12 December 1979, Atlanta, Georgia. 43 p.
2. **Goffman W & Newill V A.** Generalization of epidemic theory. *Nature* 204:225-8, 1964.
3. **Goffman W.** Stability of epidemic processes. *Nature* 210:786-7, 1966.
4. ----------------. An epidemic process in an open population. *Nature* 205:831-2, 1965.
5. **Gillis J.** The mathematical theory of epidemics. *Interdisciplin. Sci. Rev.* 4:306-14, 1979.
6. **Goffman W.** Mathematical approach to the spread of scientific ideas—the history of mast cell research. *Nature* 212:449-52, 1966.
7. ----------------. A mathematical model for analyzing the growth of a scientific discipline. *J. Ass. Comput. Mach.* 18:173-85, 1971.
8. **Goffman W & Harmon G.** Mathematical approach to the prediction of scientific discovery. *Nature* 229:103-4, 1971.
9. **Hawkins D T.** The literature of noble gas compounds. *J. Chem. Inform. Comput. Sci.* 18:190-9, 1978.
10. **Price D J D & Gürsey S.** Studies in scientometrics. Part 1. Transience and continuance in scientific authorship. *Int. Forum Inform. Doc.* 1(2):17-24, 1976.
11. ------------------------------. Studies in scientometrics. Part 2. The relationship between source author and cited author populations. *Int. Forum Inform. Doc.* 1(3):19-22, 1976.
12. **Warren K S & Goffman V.** The ecology of the medical literatures. *Amer. J. Med. Sci.* 262:267-73, 1972.
13. **Goffman W & Warren K S.** *Scientific information systems and the principle of selectivity.* New York: Praeger, 1980. 189 p.
14. ------------------------------------. An application of the Kermack-McKendrick theory to the epidemiology of schistosomiasis. *Amer. J. Trop. Med. Hyg.* 19:278-83, 1970.
15. **Garfield E.** Bradford's law and related statistical patterns. *Current Contents* (19):5-12, 12 May 1980.
16. --------------. Introducing *Journal Citation Reports*®. *Current Contents* (35):5-20, 30 August 1976.*
17. --------------. *Excerpta Medica*—abstracting the biomedical literature for the medical specialist. *Current Contents* (28):5-10, 14 July 1980.
18. --------------. Most-cited articles of the 1960s. 3. Preclinical basic research. *Current Contents* (5):5-13, 4 February 1980.
19. --------------. The 100 most-cited authors of 20th century literature. Can citation data forecast the Nobel Prize in literature? *Current Contents* (4):5-11, 28 January 1980.
20. --------------. Mapping the structure of science. *Citation indexing.* New York: Wiley, 1979. p. 98-147.
21. **Goffman W & Newill V A.** Communication and epidemic process. *Proc. Roy. Soc.* A298:316-34, 1967.

*Reprinted in: **Garfield E.** *Essays of an information scientist.* Philadelphia: ISI Press, 1980. 3 vols.