Networking the Internet: *A Proposal for the Network Analysis and Visualization of Internet Pioneer Influences*

By Anya Kalogerakos
05/07/2024

Research Question:

The most prolific network in history, the Internet, is now responsible for linking 66% of the globe's population. Despite its wide reach, when the question of Internet inventorship is posed, a surprisingly narrow list of "Internet Pioneers" are credited for their work, though countless more researchers are likely involved. Though many Internet history sites provide information on Internet Pioneers, which can give a sense of the ideology behind the Internet technology we know today, what is missing is the foundational work these pioneers built off. As one can imagine, in the 1940s era of the first programmable computer (the ENIAC), thoughts of a communicating computer system were not far off. Despite the narrative often told, the Internet Pioneers were not the first to conceive of the Internet, nor did they create this technological wonder without basis. Instead, a hidden foundational layer of research in the early to mid-1900s supported their progress—and is nearly invisible to investigators of Internet history.

An example of these missing researchers can be easily found when one knows what they are looking for. According to a brief written history of the Internet by several Internet Pioneers (including Vint Cerf and Bob Kahn), "Leonard Kleinrock at MIT published the first paper on packet switching theory in July 1961 and … convinced Roberts of the theoretical feasibility of communications using packets rather than circuits." [1] This quote describes how Kleinrock worked with Lawrence Roberts, one of the chief scientists at DARPA, to recommend novel packet-based computer communication. It also names one of the earliest authors of Internet-related research—Leonard Kleinrock. When one looks for the 1961 paper referenced in this quote (see Figure 1), beyond brilliant work, one notices 30 citations to work as old as the 1920s ranging from "Routing Procedures in Communications Networks" (Pollack, 1960) to "Introduction to Congestion in Telephone Systems" (Syski, 1960) (see Figure 2). This proves that there does exist a foundational framework, largely built off computer electronics and telephone communication networks, that inspired our modern Internet. However, this begs the question, who are the underrepresented individuals the Internet Pioneers built upon?

As such, the goal of this project will be to compile information on the researchers whose work the Internet Pioneers used within their own research. Questions on who they were, what they studied, what they contributed to the Internet, and how their work formed relationships with the more commonly known Internet Pioneers will all be investigated. The final product of this research will be a network analysis and visualization of the Internet Pioneers and their cited authors during the development of the Internet (roughly from 1960-2000). A citation network was chosen for this digital humanities project, as they possess the unique ability to "give a large-scale perspective on

[1] Cerf, Vinton, et al. "Origins of the Internet."

the history of science, identifying relational patterns across a vast number of documents that might otherwise require an entire career to digest."[2]

Identification of Dataset:

To find the work that inspired the Internet Pioneers, the Internet Pioneers themselves must first be found. There are several reliable lists of Internet Pioneers, especially those listed by Stanford's "Birth of the Internet" plaque (see Figure 3), CERN's info.cern.ch (which also happens to be the first website ever—see Figure 4), and the Internet Society's Internet Pioneers Hall of Fame.[3] With repeats, these lists give about 100 names, but they may not necessarily cover all Internet Pioneers, especially those whose work may be underrepresented or recently found to have contributed to the Internet. As such, beyond these initial, reliable websites, I would like to use the top 100 results for the Google search "Internet Pioneer" as sources and scrape any names listed in these articles, placing them in a spreadsheet for further cleaning (see Figure 5).

In this process, the main bug to consider is whether incorrect names are scrapped (for instance, the names of article authors). Realistically, some form of manual inspection will be needed to ensure the names are Internet-related. Using OpenRefine, I would like to import the .csv of Internet Pioneer names collected from each website to count the frequency of names mentioned and combine all the names into a single column (no repeats), with the frequency of mention in the adjacent column (see an example done in Google Sheets in Figure 6). Then, using this organized list, I will inspect the names that receive 10% or lower of the highest possible citation number. For instance, if Vint Cerf is cited 60 times, then names with under 6 citations will be investigated via Google search for career information. This will likely remove any improperly added names due to scrapping errors.

From here, more data will be added to this list of names and website mention frequencies to provide a more robust understanding of each pioneer for the later network analysis. Demographic information such as age during publication, nationality, geographic location at time of research, level of education at time of research, academic institution attended, place of employment at time of research as well as information on relevant research such as publication title(s), year of publication(s), and publication journal(s). Demographic information will be collected to illuminate the different types of people involved in the formation of the Internet so that patterns between the Internet Pioneers can be visualized and inform what perspectives have been behind the Internet. To collect this information, the best methodology appears to be to scrape these data fields from Wikipedia articles, which often tag age, education, nationality, etc. However, since I would like to center this information around the time that their most important contributions to the Internet were made, some adjustments may need to be made to correctly calculate age during publication, education at the time of publication, employment at the time of publication, etc. This will unfortunately have to be done manually, though there are many resources to confirm Internet Pioneer demographic information, as a vast majority keep a strong online presence and have participated in many Internet-history-preserving projects.

---

[2] Painter, Deryc, et al. "Abstract."

[3] See list of "Internet Pioneer Hall of Fame" inductees here.

The previous step highly relies on relevant publication(s) by the Internet Pioneer, the identification of which will itself be reliant on Google Scholar. Publication information will be collected in order to facilitate a citation network that connects the Internet Pioneers to those who inspired their work. Citation networks are reliably used to show the influence of different scholars on an author's work as "it is generally assumed that a citation represents the citing author's use of the cited work and indicates an influence of the cited work on the author's new work."[4] Ideally, Internet Pioneer's most relevant work to the foundation of the Internet will be collected, although this task is difficult to accomplish in practice. By using a Python library "scholarly" to automate the process of searching and retrieving information on each Internet Pioneer's name within Google Scholar, the title, year of publication, the publication journal, any citations, and a link to the relevant paper will all be collected. If there are any co-authors on the paper, they should be listed and will later be included in the Internet Pioneer network. What constitutes a "relevant paper" will be the most difficult thing to automate. In an attempt to do this, only papers between 1960 and 2000 will be considered, as that is when most Internet research was being conducted. Then, either the most highly cited publication *or* any publication cited by 1000 or more researchers will be collected. However, this method of collection is still susceptible to error and will have to be manually reviewed. A great example of why this is needed can be shown by searching Internet Pioneer "Paul Baran" on Google Scholar (see Figure 7). Before his Internet-relevant 1964 paper, two highly cited "Paul Baran" papers appear. However, these highly cited papers belong to an economist named Paul Baran, not the Internet Paul Baran in question. Therefore, each article title and publication journal should be reviewed after collection to ensure the research is correctly related to the Internet and not confused by a shared name.

Once the relevant articles are found and titles recorded, their text will need to be located (which can typically be done via Google Scholar). If the citations are not listed in the metadata provided by Google Scholar, pages that include citations and references will need to be collected and downloaded for each relevant article. Using an OCR tool like Transkribus, the citations will need to be converted into legible text with both the author and the title of each cited paper tagged as "author" and "title." All works and authors cited by a particular Internet Pioneer will then be added as another field in their corresponding .csv file row. If a citation contains "et al.," the paper cited will need to be manually searched for and author(s) recorded. This information must be formatted in a way that allows use in the creation of the network. Ideally, citation entries will occupy the same cell in a row but will be formatted as the following:
"Author1LastName_Author1FirstName" "Paper1Title"; "Author2LastName_Author2FirstName" "Paper2Title." This way, in .csv format, each cited author will be connected to an Internet Pioneer by row but still can be separated from each other by using the semicolon marker in software like OpenRefine.

While this process uses Google Scholar, there are other data collection alternatives to consider. A weakness of Google Scholar is that it may not have all the most relevant articles in its database. In order to be thorough, it would be best to manually certify that the most influential Internet Pioneer papers were found by referencing several Internet History sources (such as the written history by Vint Cerf and Bob Kahn mentioned earlier). This would ensure that any research

[4] Zhao, Dangzhi and Strotmann, Andreas. "Introduction, 1.1"

mentioned in the written history is accounted for by matching the research publication year, author, and topic with the previously collected papers from Google Scholar. Another possibility outside Google Scholar, which could prove useful as Google Scholar works are often copyrighted and information beyond metadata may be behind a paywall, would be to use U.S. patent citations. Patents are typically not copyrighted and are published for public view by the U.S. Patent and Trademark Office. For an invention of a relevant process or system by an Internet Pioneer, they must cite related patents during the patent application process. Since "patent citations are references to already existing technology within either patents or scientific literature based on which the current patent is modeled," the cited patents could be used to find researchers that Internet Pioneers built off instead of research paper citations.[5] A fundamental problem with this is that much of the early Internet technology was not patented or covered by any form of intellectual property, as the research was not meant for profit. Thus, Google Scholar appears to be the best available route.

The final step to a complete dataset will be to collect demographic information on those cited by the Internet Pioneers. Similar to what was done for the Internet Pioneers, demographic information will be scrapped from Wikipedia (if available) and added to a new .csv file, where all cited researchers (not Internet Pioneers) should be listed, along with the articles they are cited in and the Internet Pioneers that reference them. If the demographic information is not available on Wikipedia, a manual Google search will have to be used to fill in as many missing details as possible. Age during publication, nationality, geographic location at time of research, level of education at time of research, academic institution attended, and place of employment at time of research will all be collected. Research and citation data beyond the paper they are cited in will *not* be considered, as this would widen the network analysis and scope of the question at hand, which is about those who *directly* influenced the Internet Pioneers.

Methodology:

At this point, a cleaned dataset containing information about Internet Pioneers, their citations, and information on the authors of said citations should be available. For this project, I would like to use Gephi, so it is important to consider its limitations. It would be best to choose a single type of edge—directed or undirected—to use in the network, as "most network algorithms and visualizations break down when combining these two flavors of edges."[6] Therefore, to harness the computational functionality of the network using Gephi, I will be treating all paper authors as a node (whether Internet Pioneer or cited researcher) and all edges will be directed from the pioneer to whomever they cited. If Internet Pioneers are co-authors, there will be little way to distinguish this without looking specifically at the relevant papers listed within each node (as they will share the same title(s). If co-authors are cited by an Internet Pioneer, they will be split into two nodes with two different directed edges, though they will share a paper title in their node information.

The dataset should be imported into Gephi such that only node names and edges are considered. For instance, only the Internet Pioneer .csv file with the list of Pioneers and the researchers they cited should be included. All the other data will be included in an interactive version of the network. Once nodes/edges are properly appearing on Gephi, the appearance

[5] Chakraborty, Manajit, et al. "Introduction."
[6] Weingart, Scott. "The Relationships."

settings will be adjusted such that there is no overlap and the size of the node is relative to the incoming number of edges (note that this should only affect the cited researchers, placing the focus of the network onto the non-Internet-Pioneers). I would then like to run "Betweenness centrality" and "Community detection" to see if any notable patterns appear in who is being cited by Internet Pioneers. An interesting finding could be if a particular researcher is cited so frequently that they possibly deserve the title of "Internet Pioneer." Any relevant information discovered here, as well as information about data collection and methodology, can be listed in the "About" section of the interactive website that will be created for dissemination.

Presentation and Dissemination:

The ultimate end goal for this project is to have an easy-to-understand visualization of those involved in the foundations of the Internet—and especially to highlight the work the Internet Pioneers built upon. As such, I would like to make an interactive network website, similar to the Shakespeare and Co. project.[7] If possible, I would like to add a locational aspect, as well as adjust the size of node markers to match the proportion of incoming edges. To do this, I would likely use the open-source Leaflet library. In my visualization, I imagine there would be a base network, imported using the Gephi library, on top of which users could hover over nodes and edges to learn more information. Each node would have a visible name without any hovering. Once a user hovers over this node, demographic information and information about their research (previously collected in the dataset) would appear on the right-hand side. See Figure 8 and 9 for an example of this visualization. If a user were to hover over an edge, the title of the research paper that connects the two nodes would pop up, as well as a link to an accessible form of the paper provided by Google Scholar (if possible/not behind a paywall). The user would be able to apply filters as well, such as specifying the years viewed, so that only researchers who published a paper within a specified year range appear. This could help assess the network's "mechanism of growth," which is a useful feature visible within dynamic citation networks.[8]

Naturally, preserving this work would entail purchasing and maintaining a domain name. If this is not possible, the website code could be uploaded to GitHub and the GitHub pages mechanism could be used to host the website, though it could not handle high traffic. Regardless of how the website is hosted, I would like to upload all of my website and network code to a GitHub directory to encourage further work and use. Similarly, I would like to place my collected dataset(s) on Zenodo so that the dataset(s) will remain accessible to those who would find it valuable. With this network visualization project, I hope to further research on Internet history, as well as inform the general public about the voices behind the Internet. As focused on by this project, while our society grows continually more dependent on the architecture and values embedded within the Internet, it is critical we understand the influences behind the technology we have developed.

[7] See https://shakespeareandco.app/.
[8] Golosovky, Michael. "Introduction, 1.1.3"

Bibliography

Cerf, Vinton, et al. “A Brief History of the Internet.” Internet Society, 1997, https://www.internetsociety.org/internet/history-internet/brief-history-internet/. Accessed 27 April 2024.

Chakraborty, Manajit, et al. “Patent Citation Network Analysis: A Perspective from Descriptive Statistics and ERGMs.” PLOS ONE, vol. 15, no. 12, Dec. 2020, p. e0241797, doi:10.1371/journal.pone.0241797. Accessed 26 April 2024.

Golosovsky, Michael. Citation Analysis and Dynamics of Citation Networks. Springer Cham, 2019, https://link-springer-com.ezproxy.princeton.edu/book/10.1007/978-3-030-28169-4. Accessed 27 April 2024.

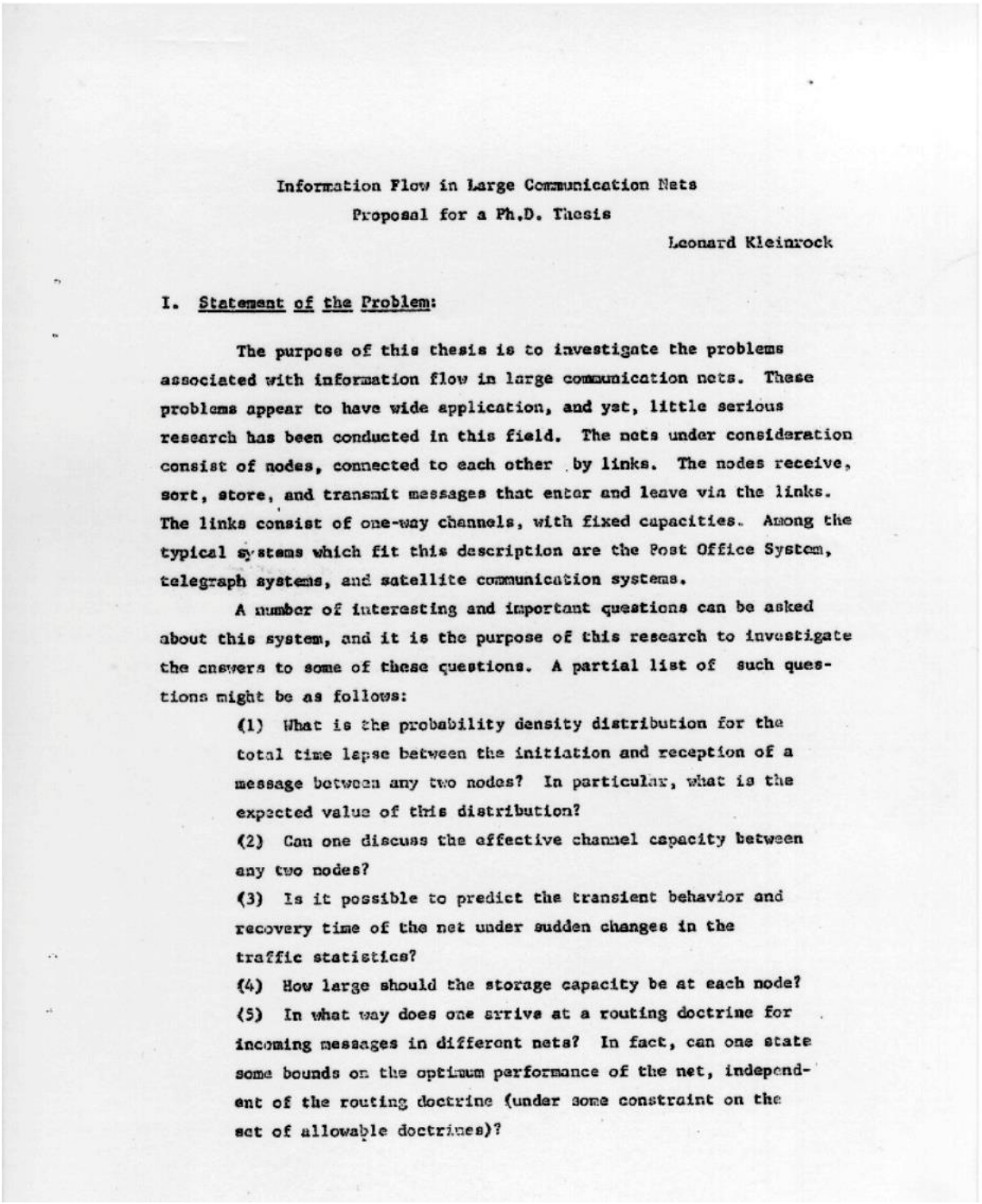
Norman, Jeremy. “Kleinrock Introduces the Concept Later Known as Packet Switching : History of Information.” Historyofinformation.Com, https://www.historyofinformation.com/detail.php?id=795. Accessed 29 May 2024.

Painter, Deryc T., et al. “Network Analysis for the Digital Humanities: Principles, Problems, Extensions.” Isis, vol. 110, no. 3, Sept. 2019, pp. 538–54, doi:10.1086/705532. Accessed 26 April 2024.

“Shakespeare and Company Project Lab.” Shakespeare and Company Project, version 1.5.7. Center for Digital Humanities, Princeton University. September 23, 2023. http://shakespeareandco.princeton.edu/lab-credits/. Accessed 29 May 2024.

Weingart, Scott B. “Demystifying Networks, Parts I & II.” Journal of Digital Humanities, vol. 1, no. 1, 2011. https://app.perusall.com/courses/introdh24/demystifying-networks-parts-i-and-ii-journal-of-digital-humanities-339819935. Accessed 27 April 2024

Zhao, Dangzhi, and Strotmann, Andreas. Analysis and Visualization of Citation Networks. Springer Cham, 2015, https://link-springer-com.ezproxy.princeton.edu/book/10.1007/978-3-031-02291-3. Accessed 27 April 2024



Information Flow in Large Communication Nets

Proposal for a Ph.D. Thesis

Leonard Kleinrock

I. Statement of the Problem:

The purpose of this thesis is to investigate the problems associated with information flow in large communication nets. These problems appear to have wide application, and yet, little serious research has been conducted in this field. The nets under consideration consist of nodes, connected to each other by links. The nodes receive, sort, store, and transmit messages that enter and leave via the links. The links consist of one-way channels, with fixed capacities. Among the typical systems which fit this description are the Post Office System, telegraph systems, and satellite communication systems.

A number of interesting and important questions can be asked about this system, and it is the purpose of this research to investigate the answers to some of these questions. A partial list of such questions might be as follows:

(1) What is the probability density distribution for the total time lapse between the initiation and reception of a message between any two nodes? In particular, what is the expected value of this distribution?

(2) Can one discuss the effective channel capacity between any two nodes?

(3) Is it possible to predict the transient behavior and recovery time of the net under sudden changes in the traffic statistics?

(4) How large should the storage capacity be at each node?

(5) In what way does one arrive at a routing doctrine for incoming messages in different nets? In fact, can one state some bounds on the optimum performance of the net, independent of the routing doctrine (under some constraint on the set of allowable doctrines)?

Figure 1 (Source: https://www.lk.cs.ucla.edu/data/files/Kleinrock/Information%20Flow%20in%20Large%20Communication%20Nets.pdf)

BIBLIOGRAPHY

1. Brockmeyer,E., H.L.Halstrom, and A.Jensen, The Life and Works of A.K.Erlang, Danish Academy of Technical Sciences, No.2 (1948)

2. O'Dell, G.F., Theoretical Principles of the Traffic Capacity of Automatic Switches, P.O.E.E.J. 13 p.209 (1920)

3. O'Dell, G.F., An Outline of the Trunking Aspect of Automatic Telephony, J.I.E.E. 65 p.185 (1927)

4. Molina,E.C., Application of the Theory of Probability to Telephone Trunking Problems, B.S.T.J. 6 p.461 (1927)

5. Molina,E.C., The Theory of Probabilities Applied to Telephone Trunking Problems, B.S.T.J. 1 p.69 (1922)

6. Fry,T.C., Probability and its Engineering Uses, (D.Van Nostrand, 1928)

7. Palm,C., Inhomogeneous Telephone Traffic in Full-Availability Groups, Ericsson Technics 5 p.3 (1937)

8. Palm,C., Analysis of the Erlang Traffic Formulae for Busy-Signal Arrangements, Ericsson Technics, No.4 p.39 (1938)

9. Feller,W., Die Grundlagen der Volterraschen Theorie des Kampfes ums Dasein in wahrscheinlichkeitstheoretischer Behandlung, Acta Biotheoretica, 5 p.11 (1939)

10. Palm,C., Intensitatsschwankungen im Fernsprechverkehr, Ericsson Technics (Stockholm) No. 44 p.1 (1943)

11. Kosten,L.M., J.R.Manning, and F.Garwood, On the Accuracy of Measurements of Probability of Loss in Telephone Systems, J.R.S.S. Ser. B 11 p.54 (1949)

12. Feller,W., An Introduction to Probability Theory and Its Applications, (John Wiley and Sons, 1950)

13. Shannon, C.E., Memory Requirements in a Telephone Exchange, B.S.T.J. 29 p.343 (1950)

14. Riordan,F.W., Telephone Traffic Time Averages, B.S.T.J. 31 p.1129 (1951)

15. Syski,R., Introduction to Congestion in Telephone Systems, (Oliver and Boyd, 1960)

16. Morse,P.M., Queues, Inventories, and Maintenance, (John Wiley and Sons, 1958)

17. Burke,P.J., The Output of a Queuing System, Operations Research 4 p.699 (1956)

Figure 2 (Source: https://www.lk.cs.ucla.edu/data/files/Kleinrock/Information%20Flow%20in%20Large%20Communication%20Nets.pdf)

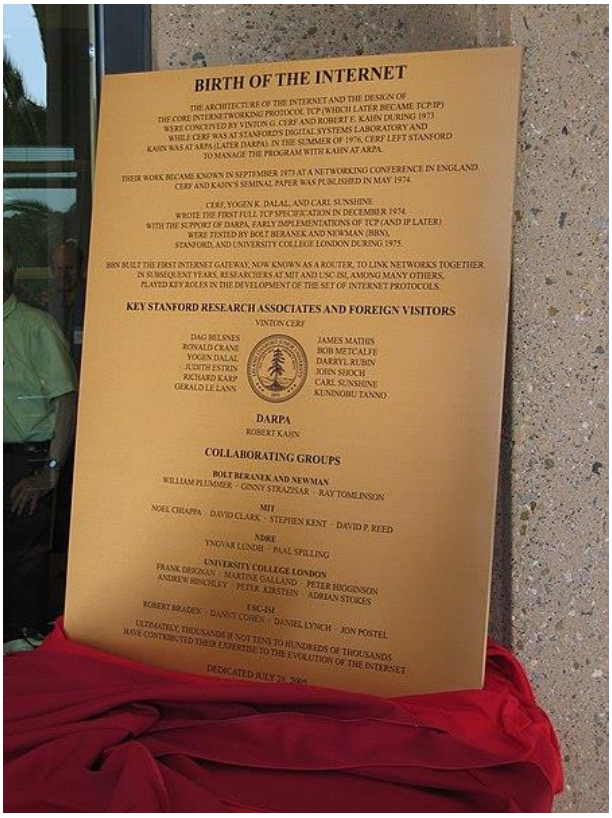


Figure 3 (Source: https://en.wikipedia.org/wiki/File:Birth_of_Internet_plaque_at_Stanford_University_July_28,_2005.jpg)

World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

What's out there?
Pointers to the world's online information, subjects , W3 servers, etc.
Help
on the browser you are using
Software Products
A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)
Technical
Details of protocols, formats, program internals etc
Bibliography
Paper documentation on W3 and references.
People
A list of some people involved in the project.
History
A summary of the history of the project.
How can I help ?
If you would like to support the web..
Getting code
Getting the code by anonymous FTP , etc.

Figure 4 (Source: https://info.cern.ch/hypertext/WWW/TheProject.html)

A42

	A	B	C	D	E	F	G	H
1	**Wikipedia**	**Stanford**	**CERN**	**ISOC**	**Website 1**	**Website 2**	**...**	
2	J.C.R. Licklider	Vint Cerf	Tim Berners-Lee	Paul Baran				
3	Paul Baran	Bob Kahn	Robert Calliau	Vint Cerf				
4	Donald Davies	Yogen Dalal	Nicola Pellow	Danny Cohen				
5	Charles M. Herzfeld	Carl Sunshine	Paul Kunz	Steve Crocker				
6	Bob Taylor	Dag Belsnes	Louise Addis	Donald Davies				
7	Larry Roberts	James Mathis		Elizabeth Feinler				
8	Leonard Kleinrock	Ronald Crane Jr.		Charles Herzfeld				
9	Bob Kahn	Bob Metacalfe		Robert Kahn				
10	Steve Crocker	Darryl Rubin		Peter Kirstein				
11	Jon Postel	Judith Estrin		Leonard Kleinrock				
12	Vint Cerf	John Shoch		John Klensin				
13	Douglas Engelbart	Richard Karp		Jon Postel				
14	John Klensin	Gerard Le Lann		Louis Pouzin				
15	Louis Pouzin	Kuninobu Tanno		Lawrence Roberts				
16	Peter Kirstein	William Plummer		David Clark				
17	Elizabeth Feinler	Ginny Strazisar		David Farber				
18	Yogen Dalal	Ray Tomlinson		Howard Frank				
19	Danny Cohen	Noel Chiappa		Kanchana Kanchanasut				
20	David J Farber	David Clark		J.C.R. Licklider				
21	Paul Mockapetris	Stephen Kent		Bob Metcalfe				
22	David Clark	David P. Reed		Jun Murai				
23	Joyce Reynolds	Yngvar Lundh		Kees Neggers				
24	Susan Estrada	Paal Spilling		Nii Quaynor				
25	Dave Mills	Frank Deignan		Glenn Ricart				
26	Radia Perlman	Martine Galland		Robert Taylor				
27	Dennis M. Jennings	Peter Higginson		Stephen Wolff				
28	Steve Wolff	Andrew Hinchley		Werner Zorn				
29	Sally Floyd	Peter Kirstein		Douglas Engelbart				
30	Van Jacobson	Adrian Stokes		Susan Estrada				
31	Ted Nelson	Robert Braden		Frank Heart				
32	Tim Berners-Lee	Danny Cohen		Dennis Jennings				
33	Robert Cailliau	Daniel Lynch		Rolf Nordhagen				
34	Nicola Pellow	Jon Postel		Radia Perlman				
35	Mark McCahill							
36	Simon Lam							

All Names | Unique Names

Figure 5

	A	B
1	**Pioneer Name**	**# of times cited**
2	J.C.R. Licklider	2
3	Paul Baran	2
4	Donald Davies	2
5	Charles Herzfeld	2
6	Robert Taylor	2
7	Lawrence Roberts	2
8	Leonard Kleinrock	2
9	Robert Kahn	3
10	Steve Crocker	2
11	...	...

Figure 6

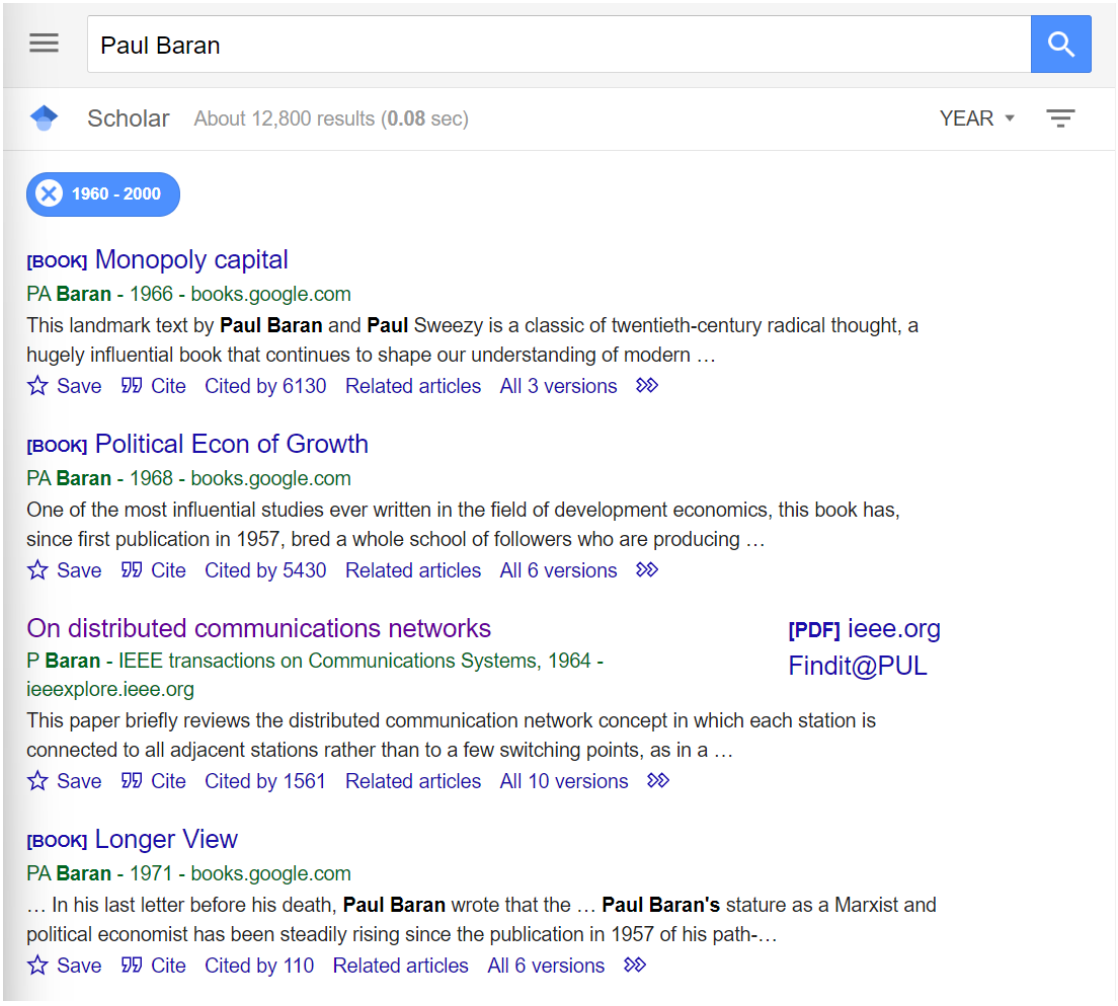


Figure 7

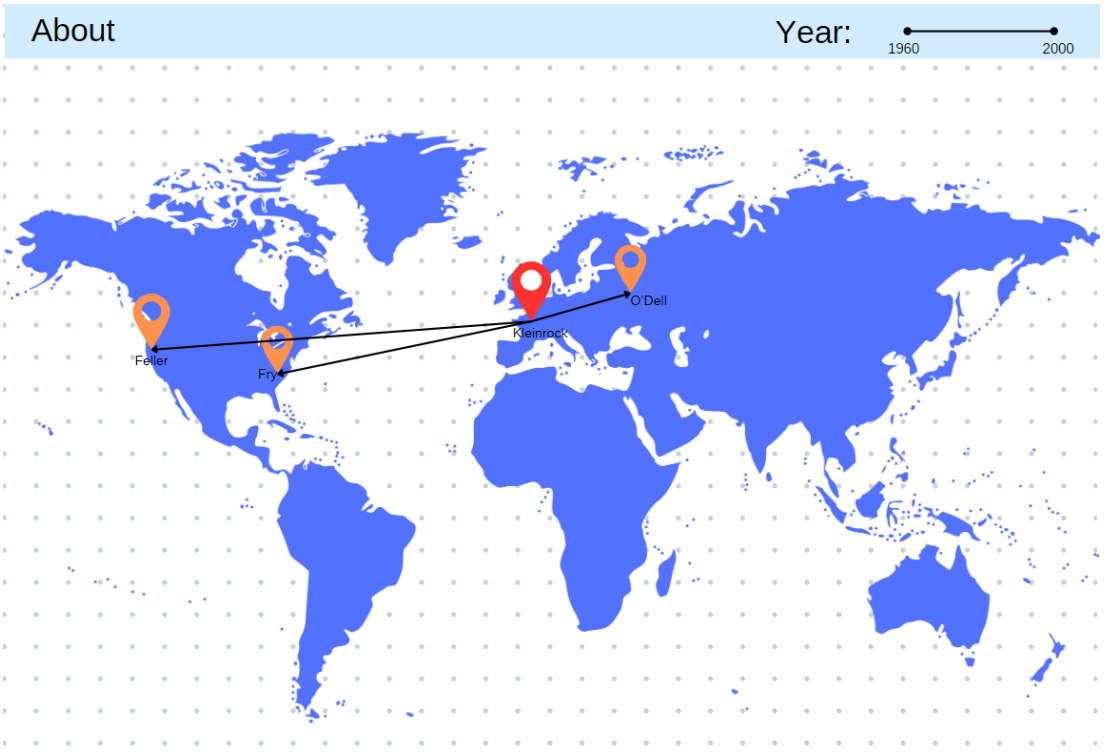


Figure 8

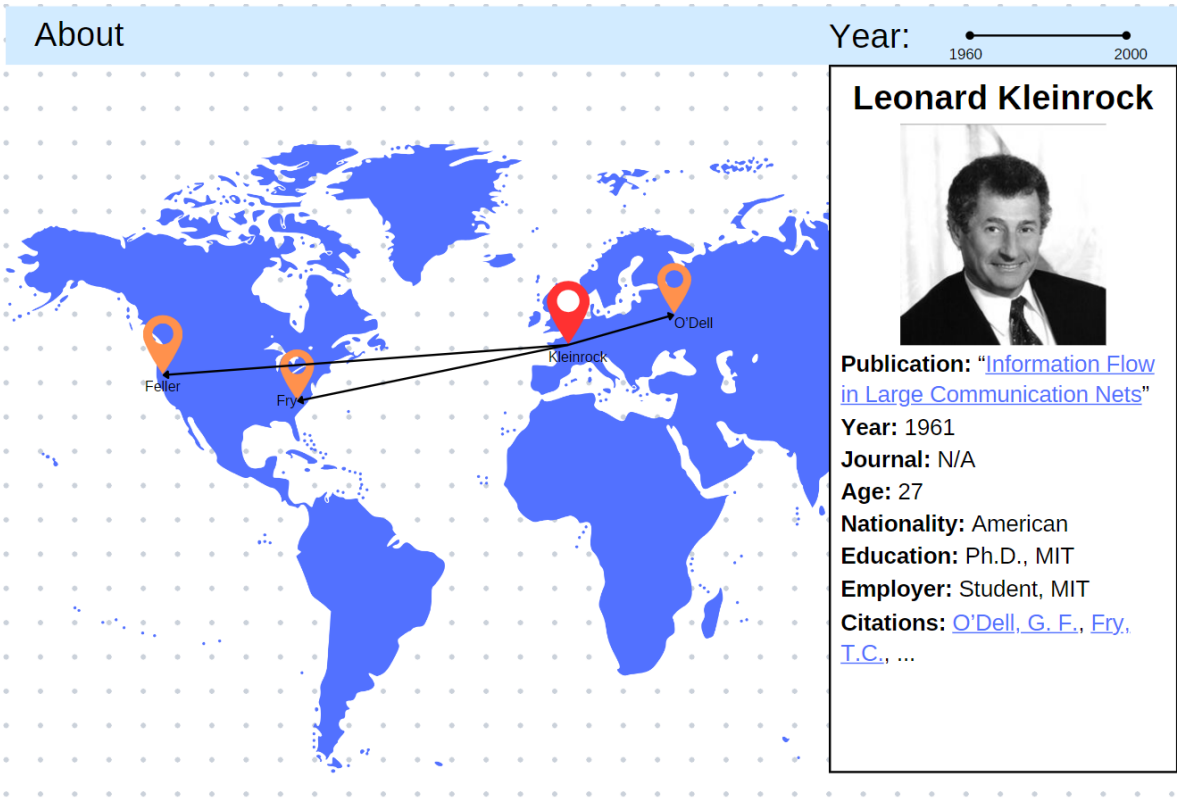


Figure 9: User hovers over Kleinrock (Note: locations are not correctly mapped in this example.)