# Mechanisms of Internet evolution & cyber risk

DISS. ETH NO. 19710

# MECHANISMS OF INTERNET EVOLUTION

# & CYBER RISK

A dissertation submitted to

**ETH ZURICH**

for the degree of

Doctor of Sciences

presented by

THOMAS-QUENTIN MAILLART
Msc. EPFL
born on January 6th 1981 in Colmar, France
Citizen of France

accepted on the recommendation of

Prof. Dr. Didier Sornette, examiner
Prof. Dr. Georg von Krogh, co-examiner
Prof. Dr. Stefan Bechtold, co-examiner

2011

# Summary

The Internet is probably the greatest communication tool ever invented yet. Most of its todays functionalities have been designed by a multitude of entities – individuals, companies, universities, governments – with no central organization. This bottom-up organization has deep implications for the evolution of the Internet itself. In this thesis, the mechanisms of Internet development are investigated, in particular individual and collective contributions to the most complex and adaptive man-made system ever achieved.

Most Internet innovations have been achieved by software development, which is made in part by original work and often by reuse of existing source code already made by others. Overall, software forms a complex directed network of modules that require other modules to work. The connectivity of this network is found to exhibit a Zipf's power law, which is a ubiquitous empirical regularity found in many natural and social systems, thought to result from proportional growth. We establish empirically the usually assumed ingredients of stochastic proportional growth models that have been previously conjectured to be at the origin of Zipf's law. For that, we use exceptionally detailed data on the evolution of open source software packages in Debian Linux distribution, which offer a remarkable example of a growing complex self-organizing adaptive system. Creation of new packages and exit of obsolete ones characterize the Schumpeterian nature of knowledge reuse in software and as a result in the development of the Internet.

The evolution of the Internet is also bounded by its interactions with the humans who shape it. Like for many technological, economic and social phenomena, the Internet is controlled by how humans organize their daily tasks in response to both endogenous and exogenous stimulations. Queueing theory is believed to provide a generic answer to account for the often observed power-law distributions of waiting times before a task is fulfilled. However, the general validity of the power law and the nature of other regimes remain unsettled. We identify the existence of several additional regimes characterizing the time required for a population of Internet users to execute a given task after receiving a message, like updating a browser. Depending on the under- or over-utilization of time by the population of users and the strength of their response to perturbations, the pure power law is found to be coextensive with an exponential regime (tasks are performed without too much delay) and with a crossover to an asymptotic plateau (some tasks are never performed). Thus, the characterization of the availability and efficiency of humans on their interactions with Internet systems is key to understand and predict its future evolution.

Among the individuals who shape the Internet, programmers are particularly important because they produce software that enables new functionalities. This work often requires having many developers cooperate to find best designs and correct mistakes. Therefore, their work is the place of intense exchange and interaction. In particular, open source software plays a crucial role in the development of new applications in a self-organized manner. Production of collective goods often requires tremendous efforts over long periods to become relevant and useful. Open source software development can be modeled as a self-excited conditional Poisson process, for which past actions, trigger – with some probability

and memory – future actions and joining of new developers. In many large – and successful – projects, these open source "epidemics" are found to be critical, hence just enough active to be sustainable.

The main drawback of self-organization is the possibility for some people to develop malicious software and use it with criminal intentions. While the Internet brings useful innovation, it is also a land of risks and uncertainty. To understand cyber risk mechanisms as a component of the Internet evolution, their statistical properties have to be established. Cyber risk exhibits a stable power-law tail distribution of damage, proxied by personal identity losses. There is also evidence for size effect, such that the largest possible losses per event grow faster-than-linearly with the size of targeted organizations.

From a risk management perspective, it would be desirable to have proper monitoring infrastructures of the Internet evolution. The Internet is a complex social world with people and organizations engaging in intense communication and thus generating numerous information transactions. Formally, these transactions can be tracked at the Internet Protocol (IP) level, like a "sniffer" on the link. If gathering, cleaning and storing data is already an issue, analyzing them is a great challenge. From an Internet security perspective, finding and characterizing security anomalies is a cumbersome work that cannot scale with terabytes of data generated by large scale networks. For that, a generalized entropy method called Traffic Entropy Spectrum (TES) has been developed and patented. It allows straightforward visual recognition of security anomalies and machine learning classification. TES is a convenient tool for real time Internet security monitoring. In the future, it could also be used for social monitoring at the IP transmission level, for instance to better assess future Internet infrastructure requirements.

# Zusammenfassung

Das Internet ist wahrscheinlich das grösste Kommunikationssystem, das jemals erfunden wurde. Die meisten der heutigen Funktionalitäten wurden durch unzählige Einzelpersonen, Unternehmen, Universitäten, Regierungen und ohne zentrale Organisation konzipiert. Dieser "bottom-up" Ansatz der Organisation hat tiefe Auswirkungen auf die evolutionäre Entwicklung des Internet selbst. In dieser Arbeit werden die zugrunde liegenden Mechanismen der Evolution des Internet und insbesondere die individuellen und kollektiven Beiträgen zu diesem komplexesten und adaptivsten vom Menschen jemals geschaffenen System untersucht.

Die meisten Innovationen im Internet wurden bei der Entwicklung von Software erreicht, welche zu einem Teil explizit neu entwickelt wurde, zum anderen aber auch durch die Wiederverwendung bereits bestehenden Quelltextes entstand. Allgemein betrachtet ist ein Programm ein komplexes Netzwerk, das aus unzähligen Modulen besteht, die wiederum selbst von anderen Modulen abhängen. Der Grad der Verknüpfungen innerhalb dieses Netzes folgt scheinbar dem Zipfschen Gesetz, das ein allgegenwärtiges empirische Gesetzmässigkeit darstellt, die in vielen natürlichen und sozialen Systemen zu beobachten ist, so als ob ein Computer Programm scheinbar ebenfalls direkt aus einem proportionalen Wachstumsprozess hervorkommen würde. In unserer Analyse verifizieren wir empirisch die allgemeinhin angenommen Elemente stochastischer proportionaler Wachstum Modelle, die bis jetzt nur als Grundlage für das Zipfsche Gesetz vermutet wurden. Dafür verwenden wir eine aussergewöhnlich detaillierte Datenbasis bestehend aus den Open Source Software-Paketen der Debian Linux-Distributionen. Diese stellen ein bemerkenswert gutes Beispiel eines wachsenden, komplexen und selbstorganisierenden, adaptiven Systems dar. Die Entwicklung neuer Paketen sowie die Einstellung der Entwicklung veralteter Pakete charakterisiert die Schumpetersche Eigenschaft der Wiederverwendung von Wissen in der Softwareentwicklung und als Folge daraus bei der Evolution des Internet.

Die Evolution des Internets ist durch die Interaktion mit dem Menschen, der es gestaltet, begrenzt. Wie für viele technologische, wirtschaftliche und soziale Phänomene, wird auch das Internet dadurch gesteuert, wie Menschen ihren Aufgaben und den Alltag als Reaktion auf sowohl endogene als auch exogene Reize organisieren. Durch die Erkenntnisse der Warteschlangentheorie wird angenommen, dass eine allgemeingültige Antwort für die oft beobachtete und einem Potenzgesetzt folgende Verteilung der Wartezeit bis zu Erledigung einer Aufgabe bereits gefunden wurde. Dennoch bleibt hier die allgemeine Gültigkeit dieses Potenzgesetzes und die Eigenschaften andere statistischer Eigenschaften nicht allumfänglich geklärt. Wir identifizieren die Existenz mehrerer zusätzlicher statistischer Eigenheiten, welche die benötigte Zeit, die Internet-Nutzern zur Abarbeitung bestimmter Aufgabe benötigen, wie beispielsweise die Aktualisierung eines Browsers, charakterisieren. Abhängig von der zeitlichen Belastung der Nutzer und deren Fähigkeit mit Störungen und Unterbrechungen umzugehen, folgt die Wartezeit einem reinem Potenzgesetz das zusammen mit einem exponentiellen Regime (Aufgaben werden hier ohne

grosse Verzögerung durchgeführt) co-existiert und einen Übergang zu einem asymptotischen Plateau (einige Aufgaben werden niemal durchgeführt) aufweist. Somit kommt der Charakterisierung von Verfügbarkeit und Effizienz der menschlichen Interaktionen mit Systemen im Internet eine Schlüsselrolle zu, die wesentlich zum Verständnis und zur Vorhersage der künftigen Evolution des Internet ist.

Von den Personen, die das Internet gestalten, sind insbesondere Programmierer von grosser Bedeutung, da diese die Software erstellen, die neue Funktionalitäten erst ermöglicht. Diese Aufgabe erfordert oft die enge Zusammenarbeit vieler Entwickler, um bessere Methoden zu entwickeln und Fehlern zu finden und zu korrigieren. Aus diesem Grund ist ihre Arbeit ein Ort intensiven Austausches und intensiver Interaktion. Insbesondere Open Source Software spielt eine entscheidende Rolle bei der Entwicklung neuer Anwendungen, welche auf eine sich selbst organisierende Art und Weise entsteht. Die Entwicklung und Produktion relevanter und nützlicher kollektiver Güter erfordert oft eine enorme Anstrengung, die über einen langen Zeitraum erbracht werden muss. Die Entwicklung von Open Source Software kann als bedingter "self-excited" Poisson-Prozess modelliert werden, bei dem die letzten Aktionen – in Abhängigkeit einer bestimmten Wahrscheinlichkeit - künftige Aktionen und das Beitreten neuer Entwickler auslöst. In vielen grossen - und erfolgreichen - Projekte, sind diese "Open Source Epidemienorganization ein sehr entscheidender Faktor, das heisst, sie sind stark genug, um nachhaltig zu sein.

Die grösste Gefahr bei selbst-organisierenden Abläufen ist die Möglichkeit, dass einige Personen schädliche Software entwickeln und gezielt mit krimineller Absicht einsetzen. Während das Internet nützliche Innovation hervorbringt, ist es auch ein Land der Risiken und Unsicherheiten. Um diese Cyber–Risiken und deren zugrundelegenden Mechanismen als eine Komponente der Evolution des Internet zu verstehen, ist es unabdingbar, ihre statistischen Eigenschaften zu verifizieren. Cyber-Risiken weisen eine stabile Schadensverteilung auf, welche einem Potenzgesetz folgt. Diese konnten wir anhand von Identitäts-Diebstählen messen und charakterisieren. Es gibt auch Hinweise auf explizite Grössen-Effekte, beispielsweise, dass der grösstmögliche anzunehmende Verlust im Schnitt schneller als proportional linear zur Grösse der attackierten Firma oder Organisation wächst.

Aus Perspektive des Risikomanagements, wäre es wünschenswert, eine geeignete Überwachungsmöglichkeit der Evolution des Internet zu Verfügung zu haben. Das Internet ist eine komplexe, soziale Welt bestehend aus Menschen und Organisationen, die in einer intensiven Kommunikation miteinander stehen und das mit der Abwicklungen von zahlreichen Informationen einhergeht. Formal können diese Transaktionen auf Ebene des Internet Protocol (IP) verfolgt werden, wie beispielsweise das"Sniffen" auf einer Verbindung. Im Fall, dass die Erfassung, Bereinigung und Speicherung von Daten bereits eine Hürde darstellt, ist es eine noch viel grössere Herausforderung, diese Daten zu analysieren. Aus der Perspektive der Internet Sicherheit ist das Aufspüren und die Charakterisierung von Anomalien eine mühsame Arbeit, die nicht einfach mit den anfallenden Terabytes in grossen Netzwerken skaliert werden kann. Zu diesem Zweck wurde eine generalisierte, Traffic Entropie Spectrum (TES) genannte Methode, entwickelt und patentiert. Diese ermöglicht

die einfache, visuelle Erkennung von Anomalien und den Einsatz maschineller Lern- und Klassifizierungsmechanismen. TES ist ein praktisches Tool zur Echtzeitüberwachung von Internet Risiken.In Zukunft könnte diese Methode auch zum Überwachen auf individueller Ebene bei der IP-Übertragung genutzt werden, um beispielsweise zukünftige Anforderungen an die Internet Infrastruktur besser einschätzen zu können.