Un modèle empirique pour la soumission de communications aux conférences

Patrick FLANDRIN

Université de Lyon, École Normale Supérieure de Lyon,
Laboratoire de Physique, UMR 5672 CNRS, 46 allée d'Italie, 69364 Lyon Cedex 07 France
flandrin@ens-lyon.fr

Résumé – On propose un modèle empirique décrivant de façon compacte le processus (cumulatif en temps) des soumissions électroniques de communications à une conférence. En amélioration au seul modèle disponible dans la littérature (basé sur un jeu unique de données), on propose un modèle de croissance global décrivant à la fois la phase d'accélération à l'approche de la date-butoir et la saturation liée au nombre nécessairement fini de soumissions. On s'appuie par ailleurs sur un corpus plus important, constitué en particulier des soumissions aux dernières conférences GRETSI. Le modèle proposé s'ajuste de façon significativement meilleure aux données observées que le seul modèle disponible jusqu'alors. Il révèle des régularités marquées et interprétables dans le processus de soumission. Il permet par ailleurs une prédiction raisonnable du nombre total de soumissions sur la base d'un nombre réduit de celles-ci.

Abstract – An empirical model is proposed for describing in a compact way the (time cumulative) process of electronic submissions to a conference. Improving upon the only available model of the literature (based on a single data set), a global growth model is pushed forward, describing both the acceleration regime when approaching the deadline and the saturation related to the necessarily finite number of submissions. The study is based on a larger data set which includes in particular the submissions to recent GRETSI symposia. The proposed model is shown to better fit the observed data as compared to the previously available one. It evidences well-defined regularities that can be given an interpretation within the submission process. It allows furthermore for a reasonable prediction of the total number of submissions on the basis of a few ones.

1 Problématique

Pour quiconque a jamais organisé une conférence, la soumission électronique des communications est un processus connu pour évoluer de façon non linéaire en temps, avec une accélération marquée à l'approche de la date-butoir, l'essentiel des soumissions se faisant dans les tout derniers jours, voire les dernières heures. Ce fait bien admis est longtemps resté qualitatif, mais un modèle (ci-après dénommé APP d'après les initiales de ses auteurs : Alfi, Parisi et Pietronero) en a récemment été proposé [1], et la question de son universalité a naturellement été posée. L'analyse proposée en [1] est basée sur une hypothèse très simple (la probabilité de soumission est supposée inversement proportionnelle au temps restant jusqu'à la date-butoir), conduisant au modèle APP qui s'ajuste remarquablement bien aux données considérées.

Deux remarques méritent cependant d'être faites à ce point :

1. La première est liée au modèle APP lui-même, qui prédit une singularité à temps fini sous la forme d'une divergence logarithmique à l'approche de la date-butoir (cf. Sect. 3.1), ce qui nécessite une régularisation dans la mesure où le nombre total de soumissions doit rester fini. En corollaire, le modèle ne permet pas de prédiction de ce nombre, nécessitant de recourir à une règle *ad hoc* indépendante [1].
2. La deuxième remarque concerne les données analysées qui se résument à une seule grande conférence (un millier de soumissions) et un atelier plus local (une grosse centaine). Il est clair qu'un corpus plus élargi est nécessaire pour rendre compte de l'universalité possible d'un modèle, celui-ci étant susceptible en outre d'être relatif éventuellement à un type de conférence et une communauté scientifique donnée.

Afin de préciser ces points, on s'attache ici à analyser les données issues de sept conférences. Ces données sont décrites en Sect. 2, leur inspection visuelle suggérant un modèle alternatif à APP (rappelé en Sect. 3.1) qui est décrit en Sect. 3.2. On s'intéresse ensuite de façon quantitative (en le comparant à APP) à l'ajustement de ce modèle en Sect. 3.3 et à son pouvoir prédictif en Sect. 3.4. On discute enfin de son interprétation et de sa possible universalité en Sect. 4.

2 Données

Le jeu de données considéré ici concerne sept conférences jouant un rôle majeur dans la communauté du signal et des images :

- 4 conférences francophones essentiellement nationales :
  - GRETSI-03 (Paris, 2003),
  - GRETSI-05 (Louvain-la-Neuve, 2005),
  - GRETSI-07 (Troyes, 2007),
  - GRETSI-09 (Dijon, 2009) ;
- 1 conférence européenne : EUSIPCO-04 (Vienne, 2004) ;
- 1 atelier international : IEEE-SSP-05 (Bordeaux, 2005) ;
- 1 exemple de la plus grande conférence mondiale du domaine : IEEE-ICASSP-06 (Toulouse, 2006).

Le détail des nombres de soumissions est donné en Table 1.

En termes d'homogénéité et de spécificité, on peut considérer que ces différentes conférences sont suivies par essentiellement la même communauté (de signal et d'image), celle-ci différant néanmoins certainement de celle (de physique statis-

TAB. 1 – Nombre effectif des soumissions pour les conférences considérées, paramètres estimés du modèle proposé (cf. Sect. 3.3) et nombre total prédit une semaine à l'avance (cf. Sect. 3.4). Dans certains cas, la date-butoir initiale a été officiellement repoussée et les résultats relatifs aux deux dates sont indiqués.

conf.	G-03	G-05	G-07	G-09	E-04	S-05	I-06
soum.	423	462	481	402	876	438	3903
β	2.6	2.5	1.6	2.5	1.6	1.3	1.7
	2	-	-	-	1	1.3	-
T (j.)	0.25	-0.1	-2	-1.3	2	1	-0.4
	-1	-	-	-	-0.2	-1.5	-
préd.	410	451	487	395	894	456	4078

tique) participant à la conférence StatPhys 23 analysée dans [1]. De plus, alors que la soumission aux rencontres StatPhys consiste à n'envoyer qu'un court résumé, la soumission aux conférences considérées ici est notablement différente dans la mesure où elle met en jeu un texte complet de 4 ou 5 pages en double colonne (ICASSP et EUSIPCO) ou un résumé étendu de 3 à 5 pages en simple colonne (SSP et GRETSI). Intuitivement, cette différence doit jouer un rôle dans le processus de soumission, les auteurs ayant une tendance naturelle à exploiter le temps qui leur est imparti avant de soumettre, et donc à le faire sans doute plus tard que s'il ne s'agissait que d'envoyer un titre et un très court résumé descriptif.

Dans tous les cas, les données ont été enregistrées de l'ouverture du site de soumission à sa fermeture, celle-ci ayant généralement lieu après la date-butoir initialement annoncée. Les instants de soumission sont connus à la seconde ("timestamps" en temps UNIX), mais une analyse les agrégeant à une échelle plus grossière (typiquement, un jour) n'a pas montré de différence notable dans les résultats. Pour cette raison, on considérera ici de telles données échantillonnées uniformément, avec de l'ordre d'une cinquantaine de valeurs dans chaque cas pour une période de soumission s'étendant sur environ deux mois.

3 Modélisation

Une inspection visuelle des différentes traces (cf. Fig. 1) révèle des ressemblances frappantes : de manière résumée, le processus de soumission (cumulé) peut être décrit par une évolution non linéaire présentant une forte accélération à l'approche de la date-butoir et une relaxation lente après celle-ci (un délai de grâce étant généralement accordé aux soumissions tardives).

3.1 Le modèle APP

À l'effet de saturation près, le modèle APP est censé rendre compte d'évolutions comme en Fig. 1. De façon plus précise, ce modèle fait l'hypothèse que la "probabilité" de soumission $p(t)$ à un instant t est inversement proportionnelle au temps restant jusqu'à la date-butoir T :

$$p(t) = \frac{C}{T-t}, \tag{1}$$

où C est une constante liée au nombre total de soumissions.

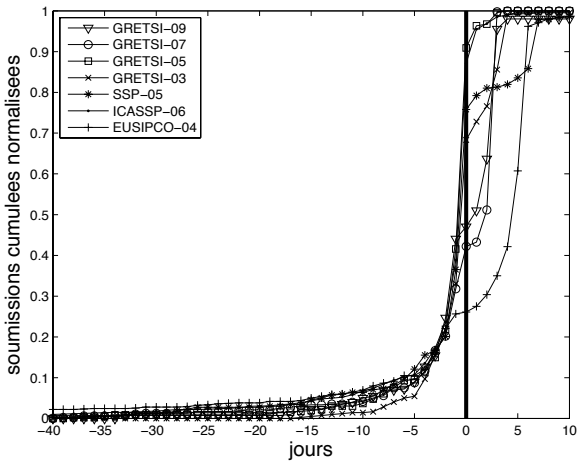

FIG. 1 – Graphes des soumissions cumulées en fonction du temps. Chaque jeu de données a été renormalisé par le nombre total de soumissions correspondant (cf. Table 1) de façon à saturer à la valeur 1, et décalé le cas échéant par rapport à la date-butoir repérée par l'instant $T = 0$ de façon à se superposer dans le régime d'accélération. Les variations après $T = 0$ correspondent à des reports de dates-butoir.

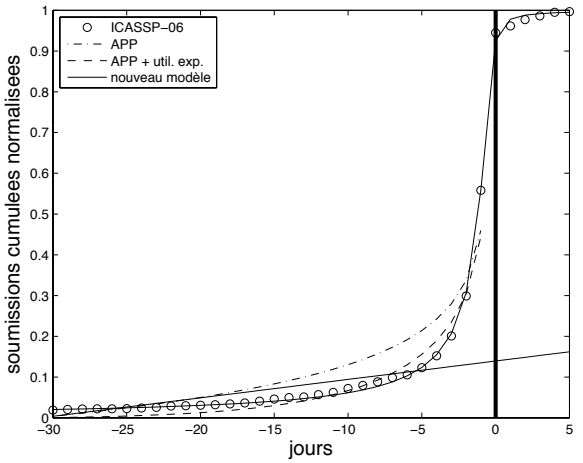

FIG. 2 – Données et ajustement de modèles au sens des moindres carrés dans le cas ICASSP-06. Le nouveau modèle, (ajusté pour tous les temps, y compris après la date-butoir repérée par $T = 0$ (trait vertical épais)) est comparé au modèle APP qui n'a de validité qu'avant la date-butoir, avec et sans fonction d'utilité exponentielle (dont l'échelle de temps estimée est ici de l'ordre de 14 jours). La droite en trait fin correspond à l'extension de l'ajustement linéaire sur la partie initiale du modèle APP, son ordonnée en $T = 0$ correspondant, après multiplication par 3, à la prédiction *ad hoc* proposée par APP pour le nombre total de soumissions.

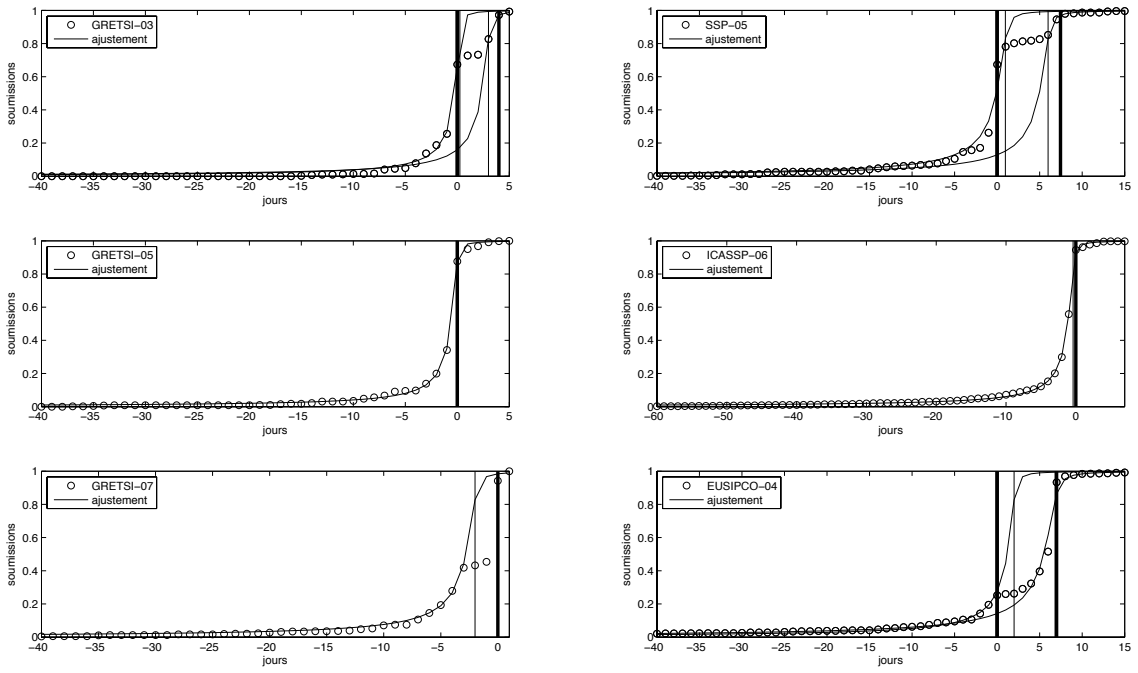

FIG. 3 – Données et ajustements aux moindres carrés du modèle (3) pour six conférences. Les dates-butoir effectives (resp. estimées) sont tracées en traits verticaux épais (resp. fins). Dans trois cas (GRETSI-03, SSP-05 and EUSIPCO-04), la date-butoir initiale (repérée ici à $T = 0$) a été officiellement repoussée et deux modèles ont été ajustés, avec une date-butoir estimée au voisinage de chacun.

La conséquence en est que le nombre total de soumissions cumulées $N(t)$ à l'instant t devrait être de la forme :

$$N(t) = \int_0^t p(s)\, ds = C\, \log\left(\frac{T}{T-t}\right), \qquad (2)$$

conduisant à une divergence logarithmique en $t = T$.

Même en ignorant le problème de la saturation qui n'est pas inclus dans le modèle APP, il apparaît que celui-ci ne permet pas de rendre compte de façon satisfaisante des données présentes (cf. Fig. 2). Une raison plausible est que les conférences considérées ici sont fréquentées par une communauté homogène mais différente de celle de la conférence StatPhys 23 considérée dans [1] pour laquelle, de plus et comme il a été rappelé plus haut, les soumissions consistent à n'envoyer qu'un court résumé.

Comme suggéré dans [1] pour les *paiements* par opposition aux *soumissions*, une "fonction d'utilité" pourrait être ajoutée au modèle de façon à prendre en compte une "pression" retardant l'action. Ceci améliore d'une certaine façon la modélisation avant la date-butoir (cf. Figure 2) mais ne règle pas le problème de la divergence à l'approche de T, ni ne permet de faire de prédiction quant au nombre total de soumissions. À ce niveau, les auteurs proposent dans [1] une règle purement *ad hoc*, à savoir ajuster une droite à la partie initiale supposée quasi-linéaire des premières soumissions et multiplier par 3 l'ordonnée obtenue à l'intersection de cette droite et de l'axe vertical à l'instant T. Cette règle est cependant sans lien direct avec le modèle et son application aux données considérées ici n'est pas concluante, cf. Fig. 2.

3.2 Un modèle empirique

Les observations et limitations précédentes conduisent à rechercher un modèle effectif dont la caractéristique principale serait une évolution non linéaire, de type sigmoïdal, comme on en rencontre dans la plupart des phénomènes de croissance qui, de façon très générale, sont souvent modélisés par des variations autour de l'équation logistique [4]. En se plaçant dans cette perspective, il s'avère que des modèles à base de sigmoïdes déformées (comme par exemple ceux dérivés de l'équation de croissance de Richards [3]) s'adaptent mal aux données alors que de meilleurs candidats (voir par exemple [2]) présentent le défaut de ne pas posséder de forme analytique.

D'un point de vue pragmatique, un des modèles possibles les plus simples pour le nombre de soumissions cumulées $N(t)$ en fonction du temps t a la forme :

$$N(t) = \frac{1 - \exp\{-\tan^{-1}[\beta(t-T)] - \pi/2\}}{1 - \exp\{-\pi\}}\, N_{tot}, \qquad (3)$$

où N_{tot} est le nombre total de soumissions et β un facteur d'échelle temporel. Afin d'accroître la flexibilité du modèle, la date-butoir T peut être considérée comme un paramètre libre, de façon à prendre en compte aussi bien une date "visée" qu'un report éventuel.

3.3 Analyse

Les graphiques de la Fig. 3 illustrent le bon caractère descriptif de ce modèle lorsqu'on en estime les deux paramètres au sens des moindres carrés (les valeurs de ceux-ci sont données en Table 1). On peut remarquer en outre que, si les va-

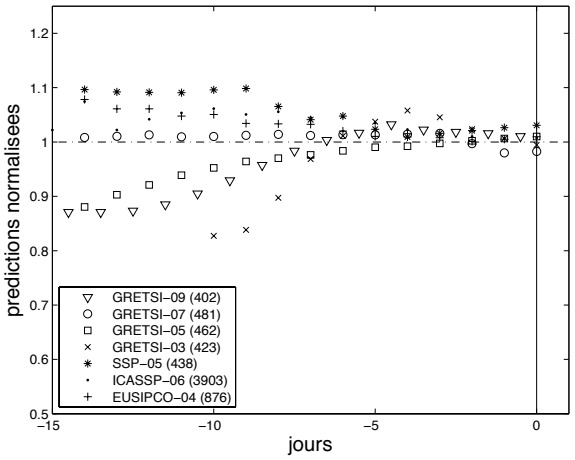

FIG. 4 – Prédictions courantes du nombre total de soumissions. Les résultats sont tracés en fonction du temps avant la date-butoir officielle (fixée à $T = 0$), avec dans chaque cas une normalisation par le nombre total des soumissions indiqué dans le cartouche. Les prédictions initiales ont été faites à un temps t_0 choisi à la moitié de l'intervalle séparant l'ouverture du site de la date-butoir, les prédictions suivantes étant obtenues par ré-actualisation glissante de la moyenne des prédictions entre t_0 et l'instant courant (sans facteur d'oubli dans le cas présent).

leurs du facteur d'échelle β présentent une certaine variabilité (entre 1 et 2.6), la même valeur ($\beta = 1.3$) est obtenue dans le cas SSP-05, avec ou sans report de la date-butoir. Des observations similaires sur GRETSI-03 et EUSIPCO-04 sont en faveur d'une interprétation selon laquelle l'annonce d'un report se traduit essentiellement par une pause dans les soumissions, sans en changer significativement le nombre total final.

3.4 Prédiction

En pratique, le nombre total de soumissions N_{tot} est un degré de liberté supplémentaire, particulièrement important pour le problème de prédiction. Étant donné la connaissance des soumissions depuis un instant initial t_0 jusqu'au temps courant t, i.e., $\{N(s|t) := N(s), t_0 \leq s \leq t\}$, l'ajustement du modèle à partir du passé disponible permet une estimation courante $N_{opt}(t)$ du nombre total de soumissions. Il est alors possible d'en déduire une estimée lissée de ce nombre selon

$$\hat{N}_{tot}(t) = \frac{1}{t - t_0} \sum_{s=t_0}^{t} N_{opt}(s); t > t_0, \qquad (4)$$

ce qui peut se faire par ré-actualisation glissante de la moyenne cumulée (avec, si nécessaire, un facteur d'oubli exponentiel).

Comme montré en Fig. 4, une prédiction très raisonnable est ainsi possible une semaine environ avant la date-butoir officielle, alors même que seulement 10% environ des soumissions sont faites.

Dans la mesure où toutes ces prédictions ont été faites sous l'hypothèse que la date-butoir effective doit se situer au voisinage de la première annoncée (aucun report de date-butoir n'est pris en compte à ce niveau), un corollaire de l'analyse conduite est que repousser la date-butoir n'accroît pas véritablement le nombre total de soumissions, mais se contente bien davantage de décaler l'avalanche des soumissions tardives, comme cela était déjà suggéré par les Figs. 1 et 3.

4 Conclusion

On a proposé un nouveau modèle pour le processus de soumission électronique à une conférence. Celui-ci diffère du premier (et seul) modèle proposé jusqu'alors [1] en ce sens qu'il inclut explicitement une saturation et rend possible une prédiction quantitative du nombre total de soumissions.

Du point de vue des organisateurs d'une conférence, les deux leçons que l'on peut tirer des analyses conduites ici sont les suivantes :

1. Ils n'ont pas à s'inquiéter trop tôt d'un faible nombre de soumissions car, en moyenne, la moitié d'entre elles sont faites le dernier jour (si l'on en croit le modèle, $N_{tot}/2$ est atteint au temps $T - 1.31/\beta$, et β est typiquement compris entre 1 et 3), et une prédiction raisonnable peut être faite une semaine plus tôt.
2. Il n'y a pas de réel avantage à repousser la date-butoir (du moins en terme de nombre de soumissions), à moins que l'on espère qu'offrir davantage de temps aux auteurs permette d'obtenir des soumissions de meilleure qualité.

La limitation de l'approche décrite ici tient bien sûr à sa nature purement phénoménologique et au fait que le modèle proposé, quoique très raisonnablement explicatif et prédictif, soit essentiellement *ad hoc*. Le modèle proposé en [1] n'était pas donné *a priori* dans la mesure où il reposait sur une hypothèse quant à la probabilité de soumission, mais cette hypothèse elle-même n'avait pas de fondement particulier (on conçoit que la probabilité de soumission augmente à l'approche de la date-butoir, mais le choix d'une proportionnalité inverse est arbitraire). Il conviendrait donc maintenant d'aborder le problème en cherchant des mécanismes génériques simples conduisant à (3) ou à des modèles analogues : c'est l'objet d'un travail en cours dont les résultats seront présentés ailleurs.

Remerciements — Merci au GRETSI, à EURASIP et à CMS pour avoir permis d'accéder aux données.

Références

[1] A. Alfi, G. Parisi and L. Pietronero, "Conference registration : how people react to a deadline," *Nature Physics* **3** (2007) 746.

[2] C.P.D. Birch, "A new generalized logistic sigmoid growth equation compared with the Richards growth equation," *Ann. Botany* **83** (1999) 713–723.

[3] F.J. Richards, "A flexible growth function for empirical use," *J. Exp. Botany* **10** (1959) 290–300.

[4] P.-F. Verhulst, "Recherches mathématiques sur la loi d'accroissement de la population," *Nouv. Mém. de l'Acad. Royale des Sci. et Belles-Lettres de Bruxelles* **18** (1845) 1-41.

[5] `www.eurasip.org`

[6] `www.gretsi.org`

[7] `www.ieee-sps.org`