Probabilistic modelisation of printer usage

Jean-Marc Andreoli - Guillaume Bouchard

Xerox Research Center Europe

Introduction

- **Objective:**
  - Analyse an printing infrastructure
  - Detect printer failures
- **Data**: Print job logs
  - At each time stamp $\mathbf{t}_1$, …, $\mathbf{t}_T$, we observe:
    - User id: $\mathbf{U}_i$
    - Printer id $\mathbf{D}_i$
- **Idea**: **Probabilistic models on (U_i, D_i, t_i)**

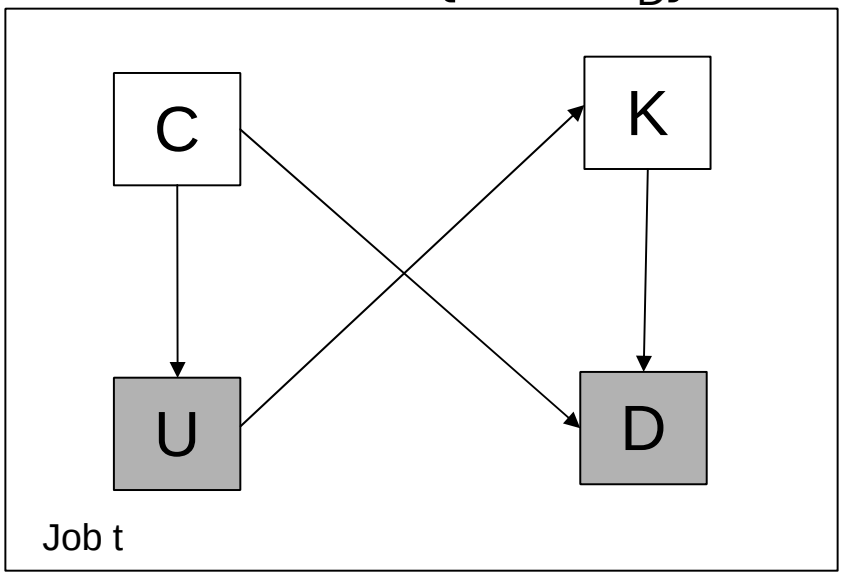
Outlines

- Co-occurrence model (static)
  - Latent class model based on user/printer co-occurrences
  - Application to XRCE printing data
- Factorial HMM model (dynamic)
  - The choice of a printer depends on the device state
  - First results on XRCE printing data

Graphical model

Random variables:

- C: job cluster in $\{1,\ldots,N_C\}$
- U: user ID in $\{1,\ldots,N_U\}$
- K: Type of job in $\{1,\ldots,N_K\}$
- D: Device ID in $\{1,\ldots,N_D\}$



Parameters (proportions):

- $\pi^{[C]}$ = P(C)
- $\pi^{[U]}$ = P(U|C)
- $\pi^{[K]}$ = P(K|U)
- $\pi^{[D]}$ = P(D|K,C)

Job category constraint

printer d cannot print category k ⇨ $\pi^{[D]}_{dk} = 0$

legend

- Observed variable
- Hidden variable

- Likelihood: mixture model

$$p(u_i, d_i \mid \theta) = \sum_c \sum_k p(u_i, d_i, c, k \mid \theta)$$

$$= \sum_c \sum_k p(u_i, d_i, c, k \mid \theta)\, p(u_i \mid \theta)\, p(c \mid u_i, \theta)\, p(k \mid u_i, \theta)\, p(d_i \mid c, k, \theta)$$

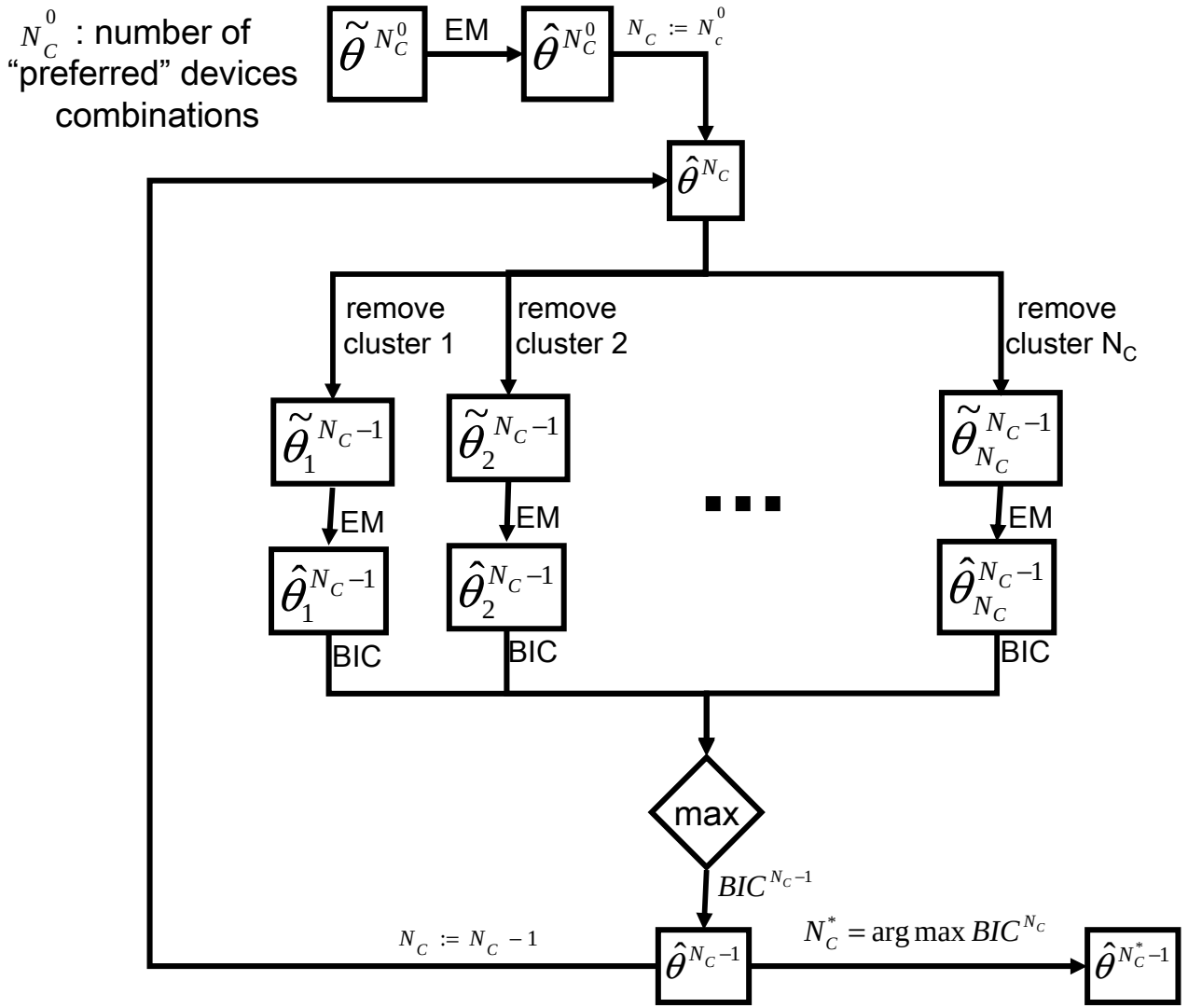
- Parameter learning: Maximum *a posteriori* (MAP) estimator

$$\hat{\theta}_{Nc} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log p(u_i, d_i \mid \theta) + \log p(\theta)$$

- Model selection: choice of the number of clusters
  - Maximal BIC criterion:

$$BIC_{Nc} = \sum_{i=1,\ldots,n} \log p(u_i, d_i \mid \hat{\theta}_{Nc}) - \frac{\upsilon_{Nc}}{2} \log(n)$$

Learning the parameters: EM algorithm



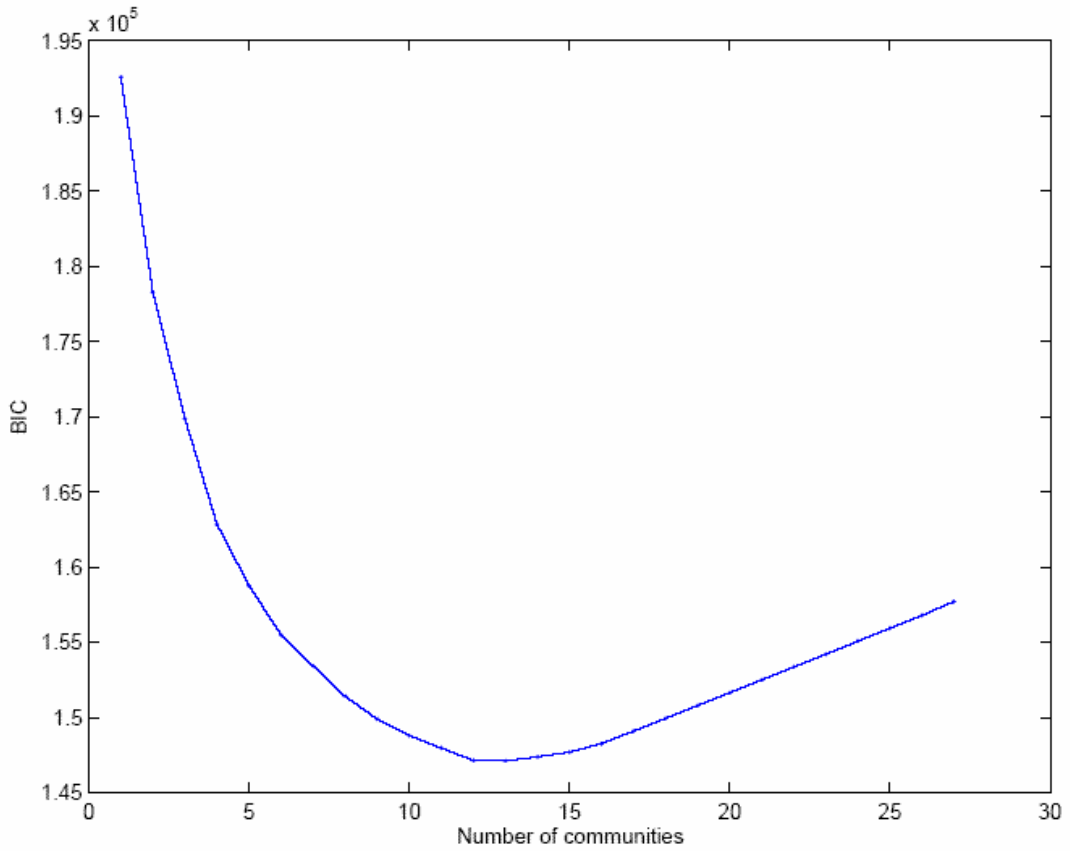
Xerox Research Centre Europe
xrce

Test on XRCE printing logs

- 27661 logs from 1/1/2004 to 05/31/2004 (120 days)
- 124 users
- 22 printers (18 B&W, 4 colour printers)

Selection of the number of clusters

Minimum of BIC: 13 clusters selected



Data sample – Matrix $A=\{n_{ud}\}$

	Tim	Vog	Pre	Stu	Hol	Bib	Her	Sta	Geo	Hod	Lem	Mes	Leq	Mid	Can	Pho	Dau	Lib	Ver	Rep	Tel	Fig
aa	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ab	149	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0
ac	12	0	0	0	2	0	0	0	902	0	0	0	0	0	0	0	0	3	67	17	0	0
ad	0	0	19	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ae	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
af	0	8	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ag	0	163	0	0	0	674	0	0	0	0	0	0	0	0	0	0	0	0	0	121	0	0
ah	0	0	27	78	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0
ai	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	32	0	0	0	0
aj	0	0	2	34	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0
ak	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
al	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0
am	0	0	69	12	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0
an	0	0	0	0	0	0	0	0	0	0	844	0	0	0	0	0	0	0	0	43	0	0
ao	0	11	64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	3	0	0
ap	0	0	0	0	0	0	0	83	0	0	0	0	0	0	0	0	0	0	0	0	11	0
aq	0	37	215	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48	1	0	0	0
ar	274	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	99	0	0	0
as	0	1	1	48	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
at	0	0	127	24	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0
au	0	0	3	153	0	0	0	0	0	0	0	0	0	0	0	0	0	56	0	0	0	0
av	0	0	0	0	0	0	32	159	0	0	0	0	0	0	0	0	0	0	0	0	36	0
aw	0	433	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	316	2	29	0	0
ax	0	0	14	130	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ay	0	0	16	0	625	0	0	0	0	0	0	0	0	0	0	0	0	28	0	0	0	0
az	150	1	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	209	0	0	0
ba	34	0	7	0	0	0	0	0	0	0	0	1	17	0	8	0	0	64	11	0	0	1
bb	0	0	0	0	0	0	86	13	0	0	0	0	0	0	0	0	0	1	0	0	54	0
bc	254	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	12	54	0	0	0
bd	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	114	0	10	3	0	0	0

Xerox Research Centre Europe

xrce

Estimated clusters

Reportage is not always the preferred color printer of Vogue users

More than 50% of the jobs

community	B&W printer	color printer	%	user IDs (% of usage)				
C1	Pre(99%)	Lib(98%)	12.7	ej(13%)	cu(9%)	bw(8%)	cm(8%)	el(8%)
C2	Stu(100%)	Lib(100%)	10	be(16%)	ds(9%)	cp(7%)	au(7%)	dc(6%)
C3	Tim(85%)	Ver(99%)	15.6	db(9%)	ar(9%)	bm(8%)	az(8%)	er(7%)
C4	Vog(99%)	Rep(52%)	13.8	cg(25%)	aw(20%)	ei(18%)	dy(15%)	ep(4%)
C5	Hol(100%)	Lib(100%)	7.7	ch(51%)	ay(31%)	bs(13%)	ec(2%)	bw(0%)
C6	Her(98%)	Tel(98%)	7	ef(26%)	dq(18%)	ce(11%)	dt(10%)	dm(8%)
C7	Geo(97%)	Ver(96%)	5.6	ac(65%)	bv(31%)	dx(2%)	eq(0%)	ec(0%)
C8	Bib(99%)	Rep(100%)	6.8	ag(42%)	bu(38%)	dh(10%)	ec(9%)	et(0%)
C9	Mes(73%)	Ver(84%)	4.5	dx(72%)	em(26%)	ba(0%)	do(0%)	bt(0%)
C10	Lem(97%)	Rep(100%)	3.5	an(92%)	ei(5%)	et(1%)	ch(0%)	bt(0%)
C11	Hod(89%)	Ver(69%)	5.5	eq(20%)	et(14%)	cy(13%)	cc(12%)	ek(9%)
C12	Mid(76%)	Fig(91%)	1.7	da(99%)	ba(0%)	do(0%)	dx(0%)	em(0%)
C13	Sta(99%)	Tel(95%)	5.6	av(12%)	de(10%)	ea(10%)	bh(8%)	cz(8%)

clusters with less than 5 users

Xerox Research Centre Europe

xrce

A more detailed view

Users assigned to cluster 1
Their printing profile is
Premiere/Libe

Community	C1: Pre/Lib
B&W printers	Pre(100%)
Color Printers	Lib(97.9%),Ver(1.6%),Rep(0.5%)
users	ej(13.2%),cu(9.6%),bw(8.9%),cm(8.9%),bt(8.2%),el(8.1%),aq(7.5%),dw(4.3%),at(4.1%),bo(2.7%),bg(2.4%), am(2.3%),ao(2.1%),cq(1.9%),ba(1.9%),ai(1.9%),es(1.7%),ee(1.6%),ct(1.3%),ad(0.5%),ae(0.5%),ak(0.4%), dp(0.3%)
Community	C2: Stu/Lib
B&W printers	Stu(100%)
Color Printers	Lib(100%)
users	be(16.9%),ds(9.6%),cp(7.8%),au(7.6%),dc(6.9%),cb(5.4%),br(5.1%),ax(4.8%),cs(3.6%),ah(3.4%),eb(2.4%), cl(2.3%),bp(2.2%),bj(2.1%),en(2.1%),eg(2%),di(1.9%),as(1.8%),aj(1.4%),bl(1.2%),dd(1.2%),dn(1.2%), cr(1%),df(0.5%),aa(0.2%)
Community	C3: Tim/Ver
B&W printers	Tim(86.2%),Leq(8.8%),Lib(4.7%),Dau(0.2%),Rep(0.1%)
Color Printers	Ver(100%)
users	db(9.9%),ar(9.4%),bm(8.3%),az(8.1%),er(7.8%),bc(7.5%),co(6.7%),bx(4.6%),ab(4%),dv(3.6%), bq(3.1%),cn(2.3%),du(2.2%),bf(2%),dz(1.9%),do(1.5%),dg(1.3%),cd(1.2%),dl(0.9%), ci(0.8%),cj(0.5%)
Community	C4: Vog/Rep
B&W printers	Vog(100%)
Color Printers	Rep(52.5%),Lib(47.5%)
users	cg(25.4%),aw(20.6%),ei(18.3%),dy(15.2%),ep(4.8%),by(1.6%),af(0.2%)

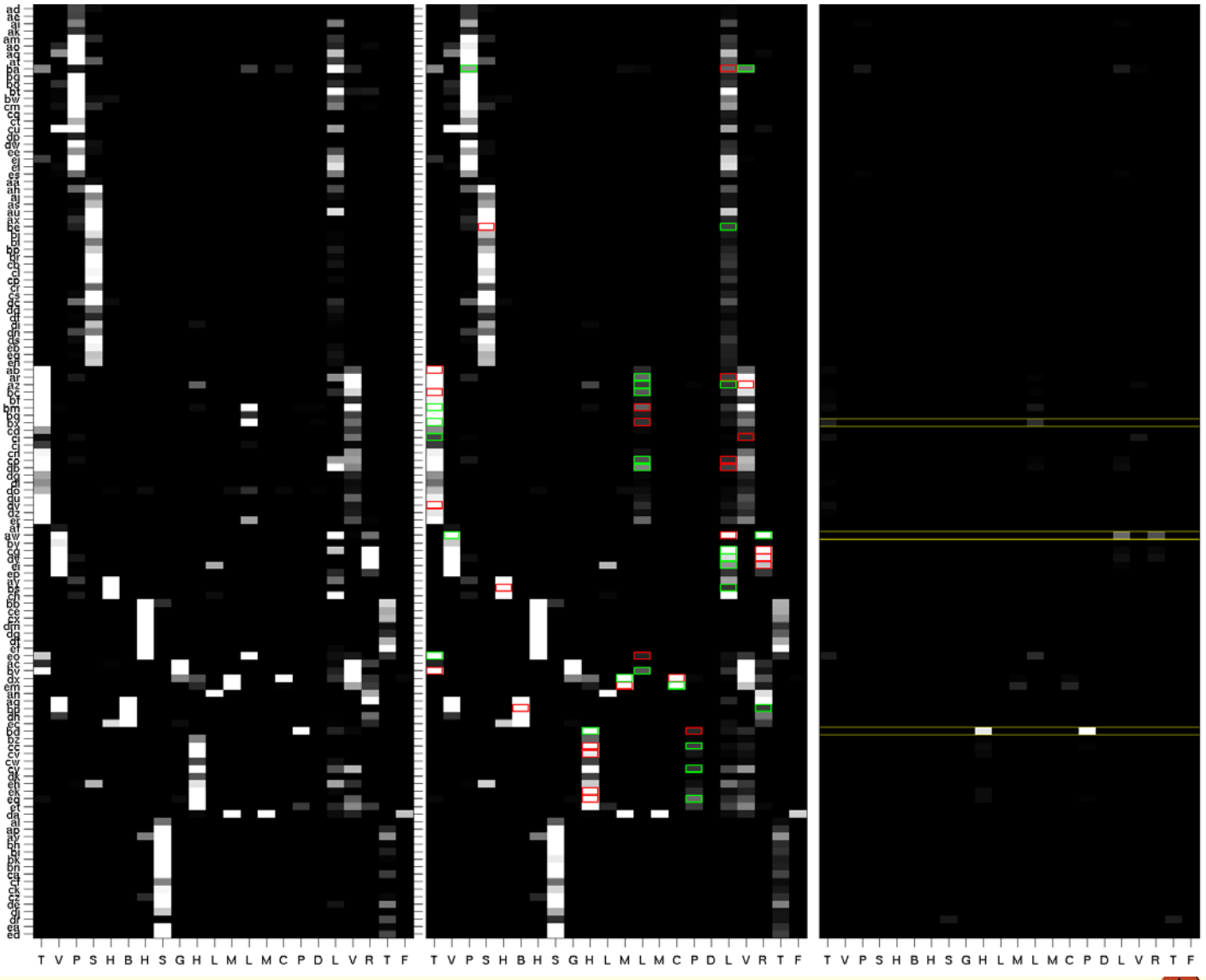
➔ Libe is also used by Vogue users

Xerox Research Centre Europe
xrce

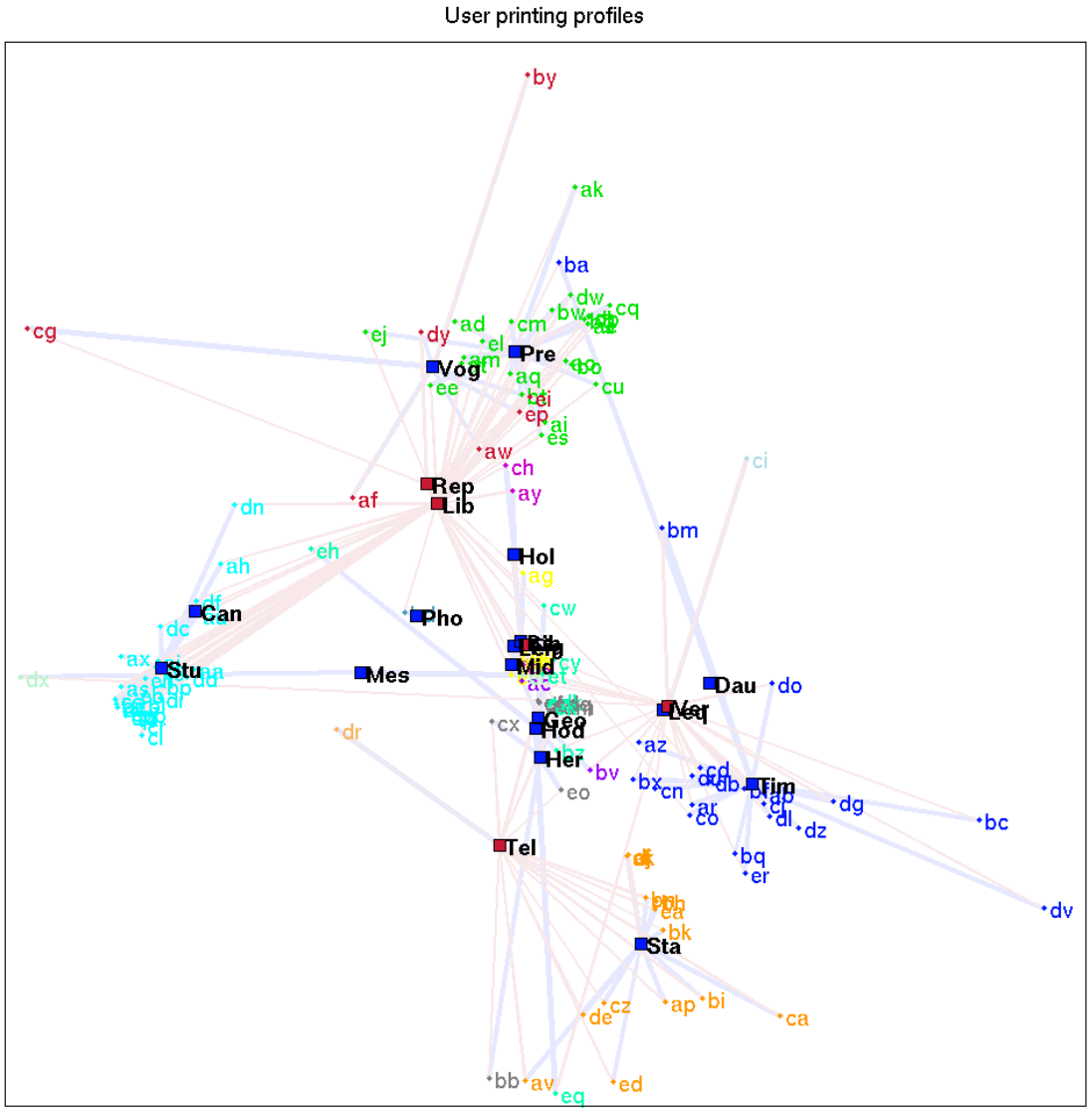
Data smoothing and outliers detection

- "smoothed" version $A^*=\{n_{ud}^*\}$ of the initial data matrix $A=\{n_{ud}\}$
  - $n_{ud}^* = n_{jobs}\, P(U=u,D=d|\pi)$
  - ➔remove undesired noise in the data (e.g. main printer can be re-estimated)
  - ➔standard methods based on A can be applied on A* with better results
- Examples:
  - Outlier detection through a Chi-squared test (at 80% level)
  - Printing infrastructure map: low-dimensional view of A and A*
  - Printer redirection estimation based on A and A*

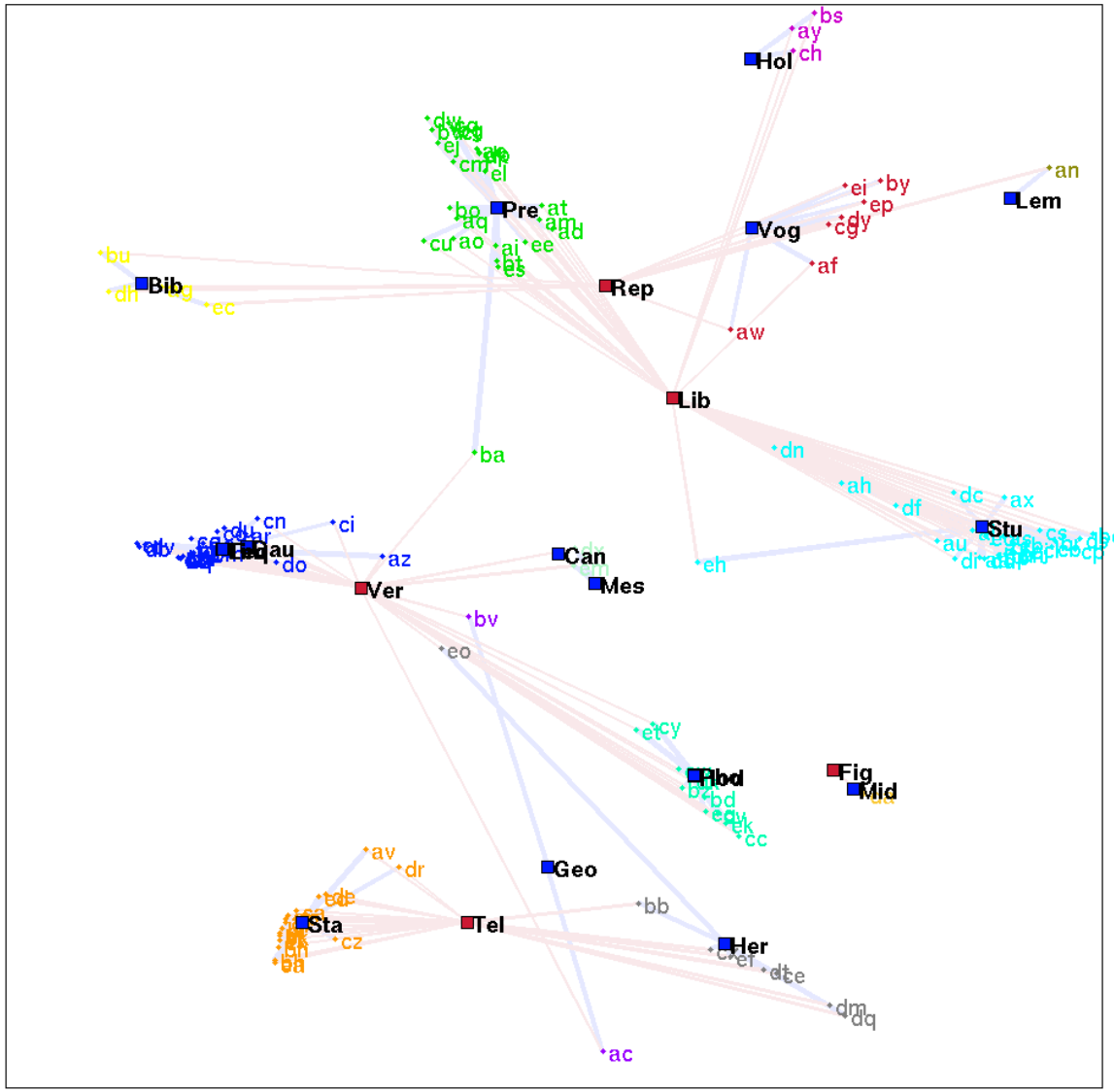
Matrices A, A* and (A-A*)^2



Infrastructure map on A - no model



Infrastructure map on A* model



Colour redirection matrices (A and A*)

Model-free color redirection matrix

	Lib	Ver	Rep	Tel	Fig
Libe	0	42	55	2	1
Vertical	74	0	21	0	4
Reportage	82	18	0	0	0
Telerama	88	12	0	0	0
Figaro	25	72	3	0	0

Model-based color redirection matrix

	Lib	Ver	Rep	Tel	Fig
Libe	0	23	75	1	0
Vertical	69	0	26	1	5
Reportage	90	10	0	0	0
Telerama	71	29	0	0	0
Figaro	19	76	5	0	0

Libe and reportage are spatially "close"

(in the same building)

Conclusion

- Probabilistic clustering model
  - fits well to the data ➔ data “denoising”
  - Simple but effective (Outliers detection, Infrastructure map, Redirection matrix)

END

B&W redirection matrices (A and A^*)

Model-free B&W redirection matrix

	Tim	Vog	Pre	Stu	Hol	Bib	Her	Sta	Geo	Hod	Lem	Mes	Leq	Mid	Can	Pho	Dau	Lib	Ver	Rep	Tel	Fig
Times	0	0	5	0	0	0	3	0	53	3	0	0	18	0	0	0	0	4	13	0	0	0
Vogue	0	0	11	0	0	57	0	0	0	0	7	0	0	0	0	0	0	10	0	14	0	0
Premiere	8	35	0	22	17	0	0	0	0	0	0	0	0	0	0	0	0	16	1	1	0	0
Studio	0	0	77	0	2	0	0	0	0	8	0	0	0	0	0	0	0	12	0	0	0	0
Hola	0	0	32	1	0	15	0	0	3	0	7	0	0	0	0	0	0	41	0	0	0	0
Biba	0	89	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0
Herald	15	0	0	0	0	0	0	20	0	0	0	0	20	0	0	0	0	1	0	0	43	0
Stampa	0	0	0	0	0	0	70	0	0	0	0	0	0	0	0	0	0	1	0	0	29	0
Geo	69	0	0	0	1	0	0	0	0	1	0	10	0	0	5	0	0	1	11	2	0	0
Hodaka	15	0	0	7	0	0	0	0	4	0	0	33	1	2	13	6	0	4	14	1	0	0
Lemonde	0	69	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0
Messager	0	0	0	0	0	0	0	0	9	7	0	0	0	12	62	0	0	1	8	1	0	1
Lequipe	78	0	0	0	0	0	15	0	0	1	0	0	0	0	0	0	1	1	3	0	0	0
Midilibre	0	0	0	0	0	0	0	0	0	3	0	84	0	0	0	0	0	1	2	0	0	10
Canard	0	0	0	0	0	0	0	0	6	4	0	85	0	0	0	0	0	1	4	0	0	0
Phosphore	12	0	0	0	0	0	0	0	0	70	0	0	4	0	0	0	0	7	6	1	0	0
Dauphine	63	0	0	0	0	0	1	0	0	1	0	1	29	0	1	0	0	0	4	0	0	0
Libe	9	36	19	4	24	0	0	0	1	2	0	1	0	0	1	0	0	0	1	1	0	0
Vertical	36	0	1	0	0	0	0	0	24	7	0	19	2	1	8	0	0	1	0	0	0	0
Reportage	0	61	1	0	0	20	0	0	4	1	9	2	0	0	1	0	0	1	0	0	0	0
Telerama	1	0	0	0	0	0	82	16	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Figaro	0	0	0	0	0	0	0	0	0	1	0	25	0	68	3	0	0	0	1	0	0	0

Model-based B&W redirection matrix

	Tim	Vog	Pre	Stu	Hol	Bib	Her	Sta	Geo	Hod	Lem	Mes	Leq	Mid	Can	Pho	Dau	Lib	Ver	Rep	Tel	Fig
Times	0	0	5	0	0	0	5	0	44	2	0	0	25	0	0	0	0	3	13	0	0	0
Vogue	0	0	11	0	0	56	0	0	0	0	8	0	0	0	0	0	0	11	0	13	0	0
Premiere	10	34	0	20	16	0	0	0	0	0	0	0	1	0	0	0	0	18	1	1	0	0
Studio	0	0	65	0	2	0	0	0	0	8	0	0	0	0	0	1	0	23	0	0	0	0
Hola	0	0	30	1	0	13	0	0	3	0	6	0	0	0	0	0	0	46	0	0	0	0
Biba	0	87	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	1	0	8	0	0
Herald	30	0	0	0	0	0	0	20	0	0	0	0	3	0	0	0	0	1	1	0	45	0
Stampa	0	0	0	0	0	0	61	0	0	0	0	0	0	0	0	0	0	1	0	0	37	0
Geo	60	0	0	0	1	0	0	0	0	1	0	11	6	0	4	0	0	1	13	2	0	0
Hodaka	9	0	0	5	0	0	0	0	3	0	3	28	1	1	9	28	0	4	9	0	0	0
Lemonde	1	64	0	0	8	0	0	0	0	4	0	0	0	0	0	0	0	1	0	22	0	0
Messager	0	0	0	0	0	0	0	0	10	9	0	0	0	12	57	1	0	0	9	1	0	1
Lequipe	77	0	2	0	0	0	2	0	13	1	0	0	0	0	0	0	0	1	4	0	0	0
Midilibre	0	0	0	0	0	0	0	0	0	3	0	83	0	0	0	0	0	1	2	0	0	11
Canard	0	0	0	0	0	0	0	0	5	4	0	85	0	0	0	0	0	0	4	1	0	0
Phosphore	3	0	0	2	0	0	0	0	1	78	1	8	0	0	3	0	0	1	3	0	0	0
Dauphine	71	0	1	0	0	0	1	0	12	1	0	0	7	0	0	0	0	1	4	0	0	0
Libe	6	35	18	7	24	2	0	0	2	2	0	0	1	0	0	0	0	0	0	2	0	0
Vertical	35	0	1	0	0	0	0	0	25	6	0	19	4	1	6	1	0	1	0	0	0	0
Reportage	0	55	1	0	0	22	0	0	3	0	11	3	0	0	1	0	0	2	0	0	0	0
Telerama	1	0	0	0	0	0	77	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Figaro	0	0	0	0	0	0	0	0	0	1	0	25	0	69	2	0	0	0	1	0	0	0

Table 6: B&W redirection matrices

Primary printer correction

B&W jobs

az	Ver → Tim
ba	Lib → Pre
bd	Pho → Hod
ci	Ver → Tim
dr	Tel → Sta
eh	Hod → Stu
es	Lib → Pre

color jobs

al	Lib → Tel
aw	Lib → Rep
ba	Lib → Ver
bd	Lib → Ver
bu	Lib → Rep
by	Lib → Rep
cc	Lib → Ver
cf	Lib → Tel
ck	Lib → Tel
co	Lib → Ver
cv	Lib → Ver
cw	Lib → Ver
db	Lib → Ver
dj	Lib → Tel
dk	Lib → Ver
ek	Lib → Ver
eo	Tel → Ver